# ESTUDIO COMPARATIVO DE LA REGULACIÓN TRANSCRIPCIONAL EN PROCESOS DE BIODEGRADACIÓN

GUILLERMO CARBAJOSA ANTONA

Universidad Autónoma de Madrid


Facultad de Ciencias

Departamento de Biología Molecular

# Estudio comparativo de la regulación transcripcional en procesos de biodegradación



Memoria presentada para optar al grado de Doctor en Ciencias por:

**Guillermo Carbajosa Antona**

Tutor: Ricardo Amils

Director: Ildefonso Cases

# SUMMARY

The control of the expression of the catabolic pathways implicated in the degradation of aromatic compounds can be exerted by a great variety of regulatory proteins. To explain this fact it has been suggested that the catabolic genes and the regulatory proteins have evolved independently. In spite of that, we have seen that in most of the cases the regulators are activators that induce the expression of the catabolic promoters in the presence of an effector molecule. This divergence of origins suggests that this is the result of convergent evolution. Besides that, induction of the catabolic genes is frequently coupled with the repression of the gene of the regulator. It has been theorized that the auto-repression of the gene of the regulator in regulatory circuits where the regulated genes are induced endows the systems with greater responsiveness, stability and robustness. Furthermore, the responsiveness and robustness is increased when the effector molecule is the first substrate of the expressed pathway, which happens to be the case when the inducer is one of the substrates of the pathway. This "logic organization" is supported by a "physical organization" where the gene of the regulator is frequently found adjacent to its regulated operon and divergently transcribed. It has been described that catabolic genes that encode enzymes implicated in biodegradation processes are frequently found in mobile elements like plasmids, transposons and genomic islands. This genetic organization not only allows for a better co-regulation but enhances the transferability of the regulatory circuit and catabolic genes as a functional unit.
We have seen that the genes that code for the same enzymatic complexes cluster together and are coded in the same operon in most of the times. Furthermore, enzymatic complexes that perform consecutive reactions are consecutively coded in the operons in the direction of transcription. This gene order can also have an influence in the transferability of complexes and connected complexes. We have seen that both more complexes and more connected complexes are transfered with this gene order than with any other random order in a horizontal gene transfer simulation experiment. Finally, it has been proposed that regulators and promoters are recruited for the control of novel pathways because they show a high degree of flexibility, "regulatory noise", that allows them to respond to new signals. We have indeed seen that the expression of the pathways can be frequently induced by compounds that are very different compared with the substrates of the pathways and that the same regulator can respond to a wide range of different compounds. Adding to this, catabolic pathways that degrade similar compounds

are under the control of regulators that belong to different families. Interestingly, when regulators are located adjacent to their regulated genes they are more conserved between homologous catabolic gene clusters, which can be seen as the selection of functional circuits by their stability and transferability properties.

# ÍNDICE

# ÍNDICE DE FIGURAS

# INTRODUCCIÓN

## Biodegradación y regulación transcripcional

Desde la revolución industrial hasta el día de hoy, se han producido y vertido en el medio ambiente numerosos tipos de compuestos aromáticos, con estructuras nuevas y diferentes a las que existían hasta el momento en la naturaleza, como consecuencia de la actividad humana. En respuesta a estas agresiones contra el medio ambiente, los microorganismos y comunidades microbianas han desarrollado la capacidad de procesar estos compuestos recalcitrantes que no forman parte de su metabolismo central, también llamados xenobióticos, por medio de transformaciones que conducen a su introducción en él (Díaz, 2004). Con cierta regularidad aparecen en la literatura científica descripciones de nuevas actividades enzimáticas descubiertas (Janssen *et al.*, 2005). Estas nuevas capacidades adquiridas por los microorganismos son de gran interés y ha habido numerosos intentos de utilizarlas para afrontar catástrofes ambientales o la polución en general (Cases y de Lorenzo, 2005a). Pero la presencia en un microorganismo de la capacidad de realizar estas transformaciones no garantiza que las vaya usar. Los microorganismos tienen un estricto control regulatorio que les permite expresar estos genes catabólicos integrando las señales recibidas en un medio cambiante (Cases y de Lorenzo, 2005b) a la fisiología de la célula (Cases y de Lorenzo, 2001). De alguna forma, la regulación transcripcional actúa como un factor limitante que restringe la producción de estas enzimas del metabolismo secundario, que transforman compuestos difíciles de degradar, permitiendo su expresión únicamente en los casos en que no hay otra fuente de carbono disponible o en respuesta a un estrés particular (Cases y de Lorenzo, 2005b). De ahí que el estudio de los sistemas regulatorios implicados en procesos de biodegradación sea de especial interés si deseamos desvelar el funcionamiento de los mecanismos que se encuentran tras el control de estos procesos y entender como los microorganismos se comportan en su ambiente natural.

# Nuevas aproximaciones al estudio de la biodegradación y herramientas disponibles

Hasta ahora las investigaciones sobre estos procesos de biodegradación se habían concentrado en el estudio de organismos modelo y su capacidad de metabolizar compuestos individuales en condiciones de laboratorio (Parales *et al.*, 2002). Sin embargo, es obvio que esto es una simplificación enorme de la complejidad de los fenómenos que ocurren en los ecosistemas naturales. A finales de los años 80, la posibilidad de crear organismos modificados genéticamente capaces de biodegradar numerosos compuestos alimentó la esperanza de poder luchar de esta manera contra la contaminación y se realizaron algunos esfuerzos destacables en esta dirección (Rojo *et al.*, 1987). Pero determinar experimentalmente los procesos que sufren todos los compuestos xenobióticos es claramente inviable (Pazos *et al.*, 2003) por lo que los métodos computacionales constituyen una alternativa más interesante cada día.

Los primeros trabajos con este nuevo enfoque llevaron a la creación de la base de datos de biocatalisis/biodegradación de Minnesota (Ellis *et al.*, 2006) en la que está contenida la información disponible sobre rutas metabólicas, reacciones y enzimas relacionadas con el catabolismo de contaminantes medioambientales y que está revisada manualmente. Contiene un sistema para predecir la ruta catabólica que recorrería un compuesto en presencia de las reacciones almacenadas en la base de datos.

El siguiente paso fue el postulado de la 'Red Global de Biodegradación' (Pazos *et al.*, 2003). Esta aproximación, por medio de la biología de sistemas, pretende desvelar propiedades que no son discernibles por medio del estudio de los componentes individuales mediante el estudio de las relaciones entre ellos (Barabasi y Albert, 1999). Tiene como premisa la consideración de la biosfera como una realidad no compartimentalizada donde los microorganismos y los compuestos difunden libremente interaccionando entre sí. Así se genera una red que describe un suprametabolismo que integra todos los compuestos biodegradables y todas las reacciones que los degradan (Pazos *et al.*, 2003).

A continuación apareció 'Metarouter' (Pazos *et al.*, 2005), una aplicación enfocada a su uso en laboratorios que trabajen en biodegradación y/o biorremediación. Permite localizar rutas de biodegradación por medio de minería de datos e integrar información de biodegradación con información genómica o de proteínas.

Finalmente, la aportación más reciente ha sido la base de datos Bionemo (Carbajosa *et al.*, 2008) que se centra en la biología molecular de los procesos de biodegradación. Esta base de datos asigna de forma manual complejos enzimáticos a las reacciones contenidas en las rutas de degradación de la base de datos de biocatalisis/ biodegradación de Minnesota (Ellis *et al.*, 2006). Integra la información metabólica con la de regulación: contiene información sobre los genes que codifican las proteínas que forman los complejos enzimáticos, los operones en los que están contenidos estos genes y la regulación de estos operones.

## Patrones generales de regulación

La regulación de rutas catabólicas actúa a dos niveles: el específico, donde un regulador induce el inicio de la transcripción de los genes catabólicos interactuando con el sustrato aromático de la ruta, o un compuesto más o menos relacionado, y el global, donde la expresión de los genes catabólicos se integra con la fisiología del huésped y, de este modo, es limitada por las condiciones de la célula (Cases *et al.*, 2005b). En esta tesis nos vamos a centrar en el estudio de la regulación específica.

## Regulación específica

Las rutas catabólicas de degradación de compuestos aromáticos no están controladas por un único tipo de reguladores si no por reguladores que pertenecen a numerosas familias distintas de reguladores que no son específicas de estos procesos de biodegradación. Entre las diferentes familias de reguladores conocidas implicadas en la degradación de compuestos aromáticos están incluidas las familias LysR , IclR, AraC/XylS, GntR, TetR, MarR, FNR, sistemas regulatorios de dos componentes (TCSs), XylR/NtrC (Díaz y Prieto, 2000; Tropel y Van der Meer, 2004) y las más recientemente añadidas LuxR (van Beilen *et al.*, 2004; Ruiz-Manzano *et al.*, 2005; Brinkrolf *et al.*, 2006; Moreno *et al.*, 2007), PadR (Gury *et al.*, 2004; Brinkrolf *et al.*, 2006) y SinR (Barragán *et al.*, 2005; Durante-Rodríguez *et al.*, 2008). Se ha propuesto la hipótesis de que los genes catabólicos y los de los reguladores han evolucionado de forma independiente para explicar este amplio espectro de tipos de reguladores (Cases y de Lorenzo, 2001). Si esto es cierto, significaría que la capacidad de responder a sustratos específicos, o compuestos relacionados, de la misma

ruta catabólica se ha desarrollado de forma independiente en numerosas ocasiones durante la evolución de este tipo de bacterias. Curiosamente, los reguladores de la familia GntR parecen ser más comunes en las bacterias Gram negativo mientras que los TCSs se encuentran con más frecuencia en las bacterias Gram positivo (Kulakov *et al.*, 2005).

Sin embargo, todas las bacterias comparten algunas características comunes que describen a todo el grupo. La mayoría de los reguladores son activadores transcripcionales, aunque la regulación por represores también es posible, que interactúan con el ADN al que se van a unir a través de un dominio de unión al a ADN de 'hélice-giro-hélice' (HTH) y que interactúan con un sustrato de la ruta catabólica que regulan, o con un compuesto más o menos relacionado (Díaz y Prieto, 2000; Tropel y Van der Meer, 2004). Una excepción a esta norma es el regulador NbzR, que controla la expresión del operón catabólico que degrada amino-fenol en *Pseudomonas putida* HS12, que posee una dominio de unión al ADN de 'cremallera de leucina' (Park *et al.*, 2001) y el recientemente descrito PadR (Gury *et al.*, 2004). Las proteínas del tipo PadR se unen al ADN por un dominio represor del tipo 'hélice alada'.

Los reguladores de rutas catabólicas de compuestos aromáticos actúan generalmente como activadores, con la excepción de aquellos que pertenecen a las familias GntR, TetR, MarR, PadR y SinR de reguladores que actúan como represores. A pesar de ello, en este caso también hay excepciones como BphR2 en *Pseudomonas pseudoalcaligenes* KF707 (Fujihara *et al.*, 2006), que es un regulador de la familia GntR que actúa como activador, y BadR en *Rhodopseudomonas palustris* que es un regulador de la familia MarR que actúa también como activador (Egland y Harwood, 1999). Por otro lado, HmgR, un regulador descubierto recientemente que controla la degradación del homogentisato en *Pseudomonas putida* U, es un represor que pertenece a la familia IclR, cuyos miembros generalmente actúan como activadores (Arias-Barrau *et al.*, 2004).

También encontramos similitudes entre familias en la organización en dominios de estas proteínas reguladoras: la mayoría de ellas tiene un dominio N-terminal que se une al ADN, mientras el dominio C-terminal contiene el dominio de recepción de señales. Como no podía ser de otra manera, aquí también encontramos excepciones: en las familias AraC y LuxR las situación es exactamente la contraria y el dominio C-terminal es que se une al ADN y el N-terminal el que recibe las señales. En el caso del componente de respuesta de los TCSs y en el de la familia XylR/NtrC la organización es la misma que en la familia AraC pero se incluye, en ambos casos, un dominio central que parece implicado en la oligomerización y en la actividad ATPasa (Díaz y Prieto, 2000;Tropel y Van der Meer, 2004).

# Circuitos de regulación

## El modelo del operón

Uno de los modelos que ha tenido más éxito a lo hora de explica la coordinación en la expresión de un conjunto de genes ha sido el modelo del operón propuesto hace casi 50 años por Jacob y Monod (Jacob *et al.*, 1960;Jacob y Monod, 1961). El modelo original, que sólo explicaba el control por represores, ha ido ampliándose, incluyendo control por activadores, expresión desde varios promotores y hasta mecanismos anti-terminación; hasta alcanzar una gran complejidad (Henkin y Yanofsky, 2002). A partir de esta idea de control coordinado de la expresión de un conjunto de genes se ha podido derivar posteriormente el concepto de circuito regulatorio (Wall *et al.*, 2004).

## Organización genómica de los elementos *cis*

Los organismos dedican una parte considerable de su ADN a codificar elementos reguladores en *cis,* y también una fracción relevante de los genes que codifican proteínas expresan factores de transcripción, tanto unos como otros implicados en controlar y coordinar la expresión de los genes y el nivel de la transcripción (Janga y Collado-Vides, 2007a). En el campo de la genómica evolutiva se ha dedicado recientemente un considerable esfuerzo para intentar entender la evolución de las regiones codificantes y su organización genómica- algo que ha estado potenciado por el aumento en la disponibilidad de secuencias de genomas completos (Marcotte *et al.*, 1999; Snel *et al.*, 2000; Tatusov *et al*, 2001). Sin embargo se ha prestado menos atención a la evolución de los circuitos regulatorios que controlan las funciones celulares. Un circuito regulatorio comprende factores de transcripción, promotores, enzimas, genes estructurales, ARNs funcionales y metabolitos (McAdams *et al.*, 2004) y puede estar detrás del control de genes del desarrollo, del ciclo celular o de rutas metabólicas. Todos estos componentes conforman una red de interacciones entre proteínas y mecanismos regulatorios genéticos que implementan una 'lógica ' basada en la bioquímica- un sistema de control- que determina como las células responden a las condiciones ambientales que detectan (Cases y de Lorenzo, 2005b). Una de las cuestiones fundamentales que se puede preguntar con respecto a estos circuitos es porqué los elementos regulatorios *cis* están organizados de determinadas formas. Podríamos hacer la asunción de que esta

organización ha sido de alguna manera re-ordenada al azar durante el transcurso de la evolución hasta disponerse de forma apropiada para cumplir los requerimientos de sistemas regulatorios individuales. Y de alguna forma, parece haber sido así. En *Escherichia coli* se ha descrito que los mismos factores de transcripción pueden actuar como activadores o represores en función de la distancia relativa de sus sitios de unión al ADN al inicio de transcripción del promotor que controlan (Collado-Vides *et al.*, 1991). Por ejemplo, el activador IlvY puede reprimir su propia síntesis uniéndose al ADN en una zona alrededor del inicio de transcripción, y activar la transcripción del promotor que controla la expresión de *IlvC* uniéndose a 60 pares de bases de distancia antes de inicio de la transcripción de este promotor (Rhee *et al.*, 1999). Estas propiedades no se limitan a los promotores asociados a sigma 70. Aunque los promotores asociados a sigma 54 están todos controlados por activadores y no requieren que estos se unan a una zona cercana para inducir la expresión del promotor (Xu y Hoover, 2001), también pueden actuar como activadores para los genes estructurales mientras reprimen su propia expresión, como es el caso de XylR (Ramos *et al.*, 1997). Posteriormente, se demostró, con una muestra mayor que la del estudio de Collado-Vides, que la acción los factores de transcripción era independiente de la familia de reguladores a la que pertenecían, reafirmándose la importancia de la distancia relativa del sitio de unión al ADN al inicio de transcripción a la hora de definir al acción que el regulador va a ejercer sobre el promotor (Madan Babu y Teichmann, 2003). Este último estudio describió que la inmensa mayoría de los activadores únicamente se unen antes del inicio de la transcripción, y que la mayoría de los represores o se unen más allá del inicio de transcripción o tienen un sitio antes y otro más allá del inicio de transcripción.

**Diseño de los circuitos**

Una implicación de esta capacidad de los reguladores de realizar diferentes acciones en función de la distancia relativa de los sitios de unión al inicio de transcripción es que les permite reprimir la expresión de su propio gen mientras activan un promotor catabólico.
¿Qué beneficios se pueden obtener de esto? Según estudios teóricos sobre el diseño de los circuitos de regulación y los principios que lo determinan podría tener que ver con la robustez, estabilidad y capacidad de respuesta del sistema (Wall *et al.*, 2004). Robustez se define como la capacidad de un sistema de permanecer inalterado ante perturbaciones externas. Estabilidad se define como la capacidad de un sistema de volver a un estado de equilibrio tras sufrir una perturbación. Finalmente, capacidad de respuesta se define

como la habilidad de un sistema de cambiar a un estado diferente tras un cambio en el medio ambiente. En estos estudios se describe como la auto-represión de los factores de transcripción incrementa la estabilidad, robustez y capacidad de respuesta de los circuitos regulatorios elementales (Savageau, 1974; Savageau, 1975; Hlavacek y Savageau, 1995; Hlavacek y Savageau, 1996; Hlavacek y Savageau, 1997; Wall *et al.*, 2003) (definiendo como circuito regulatorio elemental aquel en el que la expresión de los genes está regulada por un único factor de transcripción en respuesta a una señal bajo unas determinadas condiciones). Existen también estudios experimentales que apoyan estas teorías, específicamente la robustez (Rosenfeld *et al.*, 2002) y la estabilidad (Becskei y Serrano, 2000). También en estos estudios se explica como el tipo de acoplamiento de la expresión del regulador a la expresión del promotor afecta a la robustez, estabilidad y capacidad de respuesta. El acoplamiento es directo cuando la expresión del regulador cambia en la misma dirección que la de los genes que controla (por ejemplo, si la expresión del regulador aumenta, la de los genes regulados también), es indirecto cuando cambian en distintas direcciones y no acoplado cuando la expresión del regulador no cambia. De este modo, para circuitos controlados por activadores, el acoplamiento indirecto es el que genera una mayor capacidad de respuesta, mientras que para circuitos controlados por represores el acoplamiento directo es el que genera una mayor capacidad de respuesta. Por otra parte, la selección de un tipo particular de acoplamiento puede estar influenciada por las ventajas de maximizar la capacidad de respuesta sin renunciar a la estabilidad y robustez del sistema. Finalmente, estos estudios predicen que la auto-regulación negativa será la preferida en base a la estabilidad, robustez y capacidad de respuesta (Wall *et al.*, 2005). La presencia de auto-regulación positiva, mediada por activador, podría explicarse por la necesidad de una respuesta en la que se necesite generar una gran cantidad de enzima en presencia de una señal, ya que esta sería la única manera de conseguirlo (Hlavacek y Savageau, 1995; Hlavacek y Savageau, 1996).
Por otro lado, otros estudios teóricos predicen que los promotores que tengan una gran demanda de expresión serán seleccionados para estar regulados por activadores y los que tengan baja demanda por represores (Savageau, 1974; Savageau, 1977). También se relaciona este patrón con la capacidad de minimizar mutaciones en el sitio de unión al ADN de los sistemas controlados por represores, al estar los reguladores adheridos más tiempo al sitio de unión protegiéndolo (Shinar *et al.*, 2006).

### Organización genética de los circuitos: 'regulación de vecinos'

Esta organización lógica, en ocasiones está respaldada por una organización física. Por ejemplo, en el grupo de genes implicados en la degradación del catecol en *Pseudomonas putida*. El operón catabólico, *catBCA*, esta anexo y a 135 pares de bases del gen del regulador que controla su expresión, *catR*, y además se transcribe en la dirección contraria (Houghton *et al.*, 1995). En estudios sobre la organización cromosómica del genoma de *Escherichia coli* se ha descrito que los pares de operones que se regulan unos a otros suelen estar más cerca unos de otros que lo que sería de esperar para una red generada al azar y que, además, tienden a estar transcritos de forma divergente (Warren y Wolde, 2004). En este mismo estudio se propone que esta organización genética favorece la co-regulación al estar solapados los dominios de regulación en el ADN. Esto permite un control regulatorio adicional como la expresión correlacionada o anti-correlacionada. Otro estudio en *E. coli* describe que de 16 pares de operones con esta organización encontraron 14 que compartían al menos un sitio de unión al ADN que mediaba la co-regulación (Heshberg *et al.*, 2005). En este mismo artículo se describía que el 44% de los factores de transcripción regulan un operón adyacente al gen que los codifica en *E. coli* y e 42% en *Bacillus subtilis*. Llamaron a este fenómeno 'regulación de vecinos'. Más recientemente se ha descrito que los genes de los reguladores que responden a estímulos externos están más cerca de sus operones regulados que los que responden a estímulos internos (Janga *et al.*, 2007b).

## Conectividad

La biología molecular resulta poco práctica para tratar con grandes cantidades de información, debido a su tradicional aproximación reduccionista, y se hace necesario encontrar nuevos marcos conceptuales (Woese, 2004). Uno de los hitos en el desarrollo de estos nuevos marcos ha sido la aplicación de la teoría de redes a la problemas biológicos, en particular al estudio de la regulación transcripcional. Aplicando la teoría de redes, los nodos de un grafo representan elementos de sistemas biológicos (Barabasi y Bonabeau, 2003) que están conectados por líneas que representan relaciones entre ellos. Esta aproximación permite cuantificar la complejidad de los sistemas con parámetros numéricos, como la conectividad y la densidad. En el caso del estudio de la regulación transcripcional los nodos serían genes regulados y reguladores. Una propiedad

importante es la mencionada conectividad. La conectividad de un nodo es la cantidad de conexiones que tiene ese nodo. Como la regulación transcripcional es direccional tendremos conectividad de salida, cuantos genes regula un regulador, y conectividad de entrada, cuantos reguladores regulan un gen.

Uno de los primeros estudios en los que se aplicó la teoría de redes al estudio de la regulación transcripcional en bacterias describía la regulación transcripcional en *Escherichia coli* (Thieffry *et al.*, 1998). Allí se describía la conectividad de la red de regulación de *E. coli* donde destacaba que la mayoría de los reguladores controla un número pequeño de promotores, y de genes, y unos pocos controlan una gran cantidad, tanto de promotores como de genes (conectividad de salida). Esto llevó posteriormente a la identificación de estos últimos y a su clasificación como reguladores globales en función de variables cuantificables y no de criterios vagos como se había venido haciendo hasta ahora (Martínez-Antonio y Collado-Vides, 2003). Por otro lado, en el estudio de Thieffry *et al.* también se describía la complejidad de la regulación de los genes y promotores midiendo la cantidad de reguladores que controla cada uno (conectividad de entrada). Este tipo de análisis nos permite definir características de los circuitos de regulación de forma cuantitativa.

## Integración con la fisiología

Desde que se descubrió que la expresión de los promotores catabólicos está estrechamente conectada con la fisiología de la célula (Cases y de Lorenzo, 2001; Cases y de Lorenzo, 2005b), la comunidad científica ha hecho un esfuerzo intentando explicar los mecanismos responsables de esta integración. Por ejemplo, se ha descrito la existencia de controles post-transcripcionales implicados en controlar la carga que puede suponer la inducción gratuita de promotores catabólicos (Velázquez *et al.*, 2005; Velázquez *et al.*, 2006) y la regulación por reguladores globales como Crc (Ruiz-Manzano *et al.*,2005;Morales *et al.*,2004;Moreno *et al.*, 2007) o CRP y PTS (Cases y de Lorenzo, 1998).

Pero quizá uno de los sistemas más investigados en relación a la integración con la fisiología sea el factor sigma 54, probablemente por estar asociado a la expresión de los promotores catabólicos *Pu,* del plásmido TOL, y *Po,* del plásmido pVI150, que han sido modelos de es estudio dentro de los sistemas de biodegradación durante muchos años (Cases y de Lorenzo, 2005b). Se ha estudiado su dependencia de IHF (factor de

integración en el huésped), una proteína que, curvando el ADN, no sólo colabora con el activador si no que evita la regulación cruzada y ayuda a reclutar el complejo de la ARN-polimerasa (de Lorenzo *et al.*, 1991; Goosen *et al.*, 1995; Pérez-Martín *et al.*, 1995; Bertoni *et al.*, 1998; Sze *et al.*, 2001). También se ha propuesto un modelo para la regulación de la degradación de estireno en *Pseudomonas* en el que IHF juega un importante papel: el TCS StyRS varía los niveles de StyR-P (fosforilada), que dependen de los niveles de StyS-P, en función de la proporción relativa entre StyS-P e IHF, lo que a su vez provoca cambios en la arquitectura del promotor y modula su actividad (Leoni *et al.*, 2007).

También se ha descrito su interacción con ppGpp, la 'alarmona' (Shingler, 2003), una molécula señal que estimula de forma directa la transcripción desde los promotores *Pu* y *Po* (Carmona *et al.*, 2000). Junto con su cofactor DksA también influye en la expresión de los promotores dependientes de sigma 54 modulando la asociación competitiva con factores sigma y reduciendo la estabilidad de los complejos de la ARN-polimerasa con los factores sigma (Szalewska-Palasz *et al.*, 2007).

Finalmente, la co-regulación de los promotores *Pu* y *Po* también se ha descrito por medio de un modelo matemático que tiene en cuenta, además de otros, parámetros como el control de la fase de crecimiento de IHF, la regulación de sigma 70 durante la fase estacionaria y la contribución de (p)ppGpp tanto en la elección del factor sigma como la liberación del promotor (Van Dien y de Lorenzo, 2003). Curiosamente, únicamente incluyendo tres de estos cuatro efectos, el modelo predice que la expresión de los dos promotores estudiados se reprime durante el crecimiento exponencial y crece de forma brusca cuando las células entran en la fase estacionaria.

## Movilidad y genes catabólicos

En un principio, se pensaba que los genes catabólicos implicados en la degradación de compuestos orgánicos contaminantes estaban localizados en ADN de plásmidos en bacterias y que no se movían cuando estaban en cromosomas. Pero, en estudios detallados posteriores se han encontrado evidencias que explican parte de los mecanismos responsables del modelado de la estructura genética que permite adquirir la capacidad de metabolizar nuevos compuestos (Van der Meer y Sentchilo, 2003). Un ejemplo de esto es la estructura de mosaico que conduce a la degradación del 2,4-dinitrotolueno. Los genes implicados en la degradación de este compuesto tienen

orígenes distintos y su organización genética sugiere una progresión hacia una estructura compacta que codifique la ruta completa. En esa progresión se encuentran remanentes del ensamblaje que no participan en el catabolismo de la nueva ruta (Johnson *et al.*, 2002). En otro estudio se describe la transferencia de una región completamente idéntica entre dos bacterias en un suelo contaminado, lo que permite que esa región se incorpore a su nuevo huésped formando una nueva ruta de degradación de clorobenceno (Müller *et al.*, 2003).

Las bacterias son capaces de incorporar material genético por diferentes procesos como la transformación o la conjugación. Los procesos posteriores de recombinación homóloga y reparación del ADN pueden limitar la transferencia de ADN haciendo que sólo ocurra entre bacterias similares. Sin embargo si los genes están incluidos en plásmidos de 'gran rango de huésped' pueden extenderse sin necesidad de recombinación (Thomas y Nielsen, 2005). Se ha descrito que esta transferencia de genes entre bacterias se hace por medio de elementos genéticos móviles (MGEs) lo que permite esquivar estas limitaciones (Top y Springael, 2003). Estos MGEs incluyen plásmidos, transposones de tipo I y II y las 'islas genómicas' (Springael y Top, 2004). Un ejemplo de 'isla genómica' que contiene genes que codifican para la degradación de xenobióticos es el elemento *clc*, que es el responsable de la anteriormente mencionada transferencia del fragmento de ADN que dio lugar a la formación de una ruta de degradación del clorobenceno. Este elemento *clc* tiene 105 kb y lleva el grupo de genes *clcRABD* que codifica la mineralización de clorocatecoles por medio de la llamada ruta del corte 'orto' que fue identificada en la bacteria que degrada 3-clorobenzoato *Pseudomonas sp.* B13 (Van der Meer y Sentchilo, 2003). El elemento clc puede auto-transferirse por conjugación entre β y Y-proteobacterias y se ha demostrado que se auto-transfiere de forma eficiente en diversos ecosistemas. Se integra en la bacteria que lo recibe por medio de una integrasa incluida en el elemento, IntB13.

En estudios recientes en microcosmos se ha mostrado la capacidad de estos MGEs de distribuir una ruta catabólica por toda una comunidad microbiana adaptándola de esta forma frente a un estrés contaminante. Se introdujeron bacterias portadoras del MGE catabólico y posteriormente se introdujo en el medio el compuesto xenobiótico específico de ese MGE. En muchos casos, se observó la transferencia del MGE a miembros autóctonos de la comunidad; en algunos de los casos se produjeron cambios en la composición bacteriana de la comunidad y la degradación del compuesto xenobiótico aumentó debido a la acción de la comunidad adaptada (Springael y Top, 2004).

# Orden de los genes

Para comparar el orden de los genes entre diferentes genomas lo primero que se requiere es identificar genes equivalentes entre ellos. En esto contexto, los genes equivalentes serán los genes ortólogos. Dos genes son ortólogos si pertenecen a genomas de diferentes especies y han evolucionado de forma independiente de un mismo gen que existía en el último ancestro común a las dos especies. Por el contrario, dos genes son parálogos si se han generado a partir de un ancestro común por duplicación. La detección, o predicción, de genes ortólogos de forma sistemática y la disponibilidad de genomas completos han permitido estudiar la conservación del orden de los genes en los organismos. Los primeros a nivel genómico sobre el orden de los genes en bacterias indicaron que los ortólogos aparecen pocos menos que distribuidos al azar indicando que el orden se pierde fácilmente (Mushegian y Koonin, 1996; Watanabe *et al.*, 1997).

Pero el orden de los genes no es sólo inestable a nivel genómico si no también al nivel de los operones (Watanabe *et al.*, 1997). El porcentaje de operones idénticos entre *E. coli* y *H. influenzae*, que son dos organismos que han divergido recientemente es del 56%. El porcentaje se reduce al 13% si comparamos *E. coli* con *Helicobacter pilori*. Esto parece indicar que la destrucción de las estructuras de los operones es neutral en cuanto a la selección, al menos en la evolución a largo plazo (Itoh *et al.*, 1997). El resultado de esta falta de estabilidad en las estructuras de los operones es que hay muy pocos operones que estén conservados en muchas especies y muchos operones que están conservados en pocas (Ermolaeva *et al.* , 2001).

A pesar de que el orden de los genes a lo largo del cromosoma se pierde rápidamente, y las estructuras de los operones son inestables en la evolución a largo plazo, existe otro modo de conservación, conocida como 'conservación de los genes vecinos' (Lathe *et al.*, 2000). La 'conservación de los genes vecinos' se refiere al hecho de que, en algunos casos, las reorganizaciones genómicas, aunque alteran el orden de los genes, mantienen el 'vecindario' de genes en términos de clases funcionales y características regulatorias de los genes que son 'vecinos'. Así se define a un grupo de estos genes como 'uber-operón' (Lathe *et al.*, 2000). Como una extensión de este concepto se ha propuesto el concepto de 'vecindario de genes conectados' (Rogozin *et al.*, 2002), que se construyen localizando pares de genes vecinos conservados y agrupando posteriormente estos pares de genes. Los 'vecindarios' más conectados son los que comparten una función común.

# Formación de operones

Se han propuesto tres razones para explicar la conservación observada de los operones en bacterias: primera, divergencia reciente; segunda, transferencia horizontal de genes reciente; y tercera, fuertes restricciones regulatorias yo estructurales que seleccionan en contra la reorganización de los operones o grupos de genes conservados (Tamames, 2001;Tamames *et al*, 2001). Se han propuesto varios modelos que tratan de explicar estas restricciones que podría dirigir el origen y mantenimiento de los operones (Lawrence y Roth, 1996). Es un tema muy controvertido y todos los modelos tienen sus limitaciones e inconsistencias. Por ello, parece probable que diferentes operones pueden haber aparecido bajo diferentes fuerzas selectivas, y que la selección natural actúa de forma diferente en la generación y mantenimiento de distintos operones.

El 'modelo de co-regulación' se basa en el hecho de que la expresión en los operones está coordinada, lo que supone un beneficio para ciertos fenotipos. Estas fuerzas pueden ser importantes para el mantenimiento de los operones, pero difícilmente explican la formación ya que no se puede justificar de que forma se creó la organización genética que puso a los genes juntos y, por otro lado, la creación de las zonas reguladoras, todo ello en ausencia de una selección positiva que sólo llegaría cuando todo el sistema estuviera organizado por completo.

El 'modelo Natal' propone que los genes se originaron por duplicaciones y posteriores mutaciones para diferenciarse, en lugar de agrupar genes preexistentes, con lo que evita la inconsistencia del modelo anterior no teniendo que justificar como llegaron los genes a estar juntos. Este modelo puede explicar la formación de algunos operones, pero el hecho es que la mayoría de los operones bacterianos no presenta homología entre sus genes.

Otros modelos proponen que la mera cercanía de algunos genes en el cromosoma podría ser ventajosa y que, por lo tanto, esa sería la principal fuerza detrás de la generación de operones, con un determinado número de pasos intermedios. Los genes se irían acercando en el genoma hasta agruparse en una unidad transcripcional. Estos modelos se dividen en tres categorías: primera, los que consideran los beneficios de transcribir y traducir los genes agrupados en una zona localizada de la célula; segundo, los que apuntan hacia la amplificación de los genes agrupados como el mecanismo primitivo de co-regulación; y tercero, los que tienen en cuenta el efecto de la proximidad de los genes en la frecuencia de co-transferencia y recombinación.

En la primera categoría estaría el 'modelo de molaridad' que sugiere que el agrupamiento de los genes termina, tras la transcripción y traducción, permitiendo la concentración local de los productos génicos de los genes agrupados. Esto facilitaría la participación de los productos génicos en series de rutas catabólicas o aceleraría la formación de complejos de proteínas. La mejor adaptación permitiría una selección positiva de esta organización genética que desembocaría en la creación y mantenimiento de operones. Algunos argumentos a favor de esta teoría son el hecho de que pueden ser necesarias interacciones entre proteínas para proteger proteínas inestables o para guiarlas en su plegamiento. También se han descrito proteínas que se pliegan dependiendo unas de otras (Thanaraj y Argos, 1996).

En la segunda categoría estaría el 'modelo de amplificación' que asume que si un bloque de genes relacionados funcionalmente son frecuentemente amplificados en respuesta a la selección de aumentar el número de genes, su agrupamiento será beneficioso y positivamente seleccionable, porque provee de un mecanismo de co-regulación.

Finalmente, el 'modelo de co-adaptación de Fisher', en la tercera categoría, está basado en el concepto de coevolución. Los genes cuyos productos estén implicados en la misma ruta catabólica, especialmente si son parte de un complejo de proteínas, se presupone que evolucionan de forma coordinada. La proximidad física de estos genes puede ser seleccionada porque reduce la frecuencia de recombinación que separan *loci* coadaptados. Por otro lado, el 'modelo de la reparación multigénica' propone que, en poblaciones con bajo presión selectiva, los genes pueden ir acumulando mutaciones en numerosos genes no esenciales que estén involucrados en una ruta o que codifiquen un complejo. La ruta, o el complejo, pueden ser reemplazados por transferencia horizontal de genes y un cambio de presión selectiva pero sólo si los genes son transferidos juntos. Otro punto de vista acerca de este último modelo es el 'modelo del operón egoísta', que sugiere que el agrupamiento en genes es beneficioso para el operón pero no necesariamente para el organismo en el que está embebido. Los grupos de genes se transfieren tanto vertical como horizontalmente mientras que los genes sin agrupar sólo lo hacen verticalmente. Si la combinación de los genes que componen el agrupamiento confiere un fenotipo seleccionable, el grupo de genes, o el operón, se propagará con éxito. Nuevos genes pueden ir siendo reclutados junto al grupo si la nueva combinación es seleccionables. este tipo de dinámica es la que siguen por ejemplo las 'islas de patogenicidad' (Hacker *et al.*, 1997).

El hecho de que la contribución de la HGT a la evolución procariota sea mucho mayor de lo que se creía (Gogarten *et al.*, 1997) apoya la validez del 'modelo del operón egoísta'.

La principal inconsistencia de este modelo es que no explica como se originaron los grupos de genes esenciales, ya que la selección positiva sólo es posible si la ausencia de función no es deletérea. Se ha propuesto que esos grupos de genes se ensamblaran antes de que la vida tal y como la conocemos divergiera. Otra explicación posible, que también se ha sugerido y se aplica a operones que codifican genes que interaccionan, es que la creación de estos grupos de genes siga el 'modelo de Fisher'.

# Respuesta a nuevas señales, especificidad y coevolución entre regulación y metabolismo

### Ruido regulatorio

En un estudio reciente sobre NahR, un regulador de la familia LysR que está presente en *Pseudomonas sp.,* se realizaron mutaciones en el dominio de reconocimiento del sustrato que siempre desembocaron en la ampliación del rango de reconocimiento de sustratos del regulador (Park *et al.*, 2005). Esta capacidad de los reguladores de responder a un amplio rango de compuestos se ha descrito previamente, y se ha llamado a este fenómeno ' ruido regulatorio'. Se ha sugerido que puede permitir a los sistemas evolucionar y adquirir la capacidad de responder a nuevas señales ambientales: 'Los sistemas de control transcripcional desarrollan la capacidad de responder a nuevas señales gracias a la falta de especificidad de los promotores y reguladores preexistentes. Cuando se hace necesario, estos se irán haciendo más específicos suprimiendo las señales indeseables y refinando el ajuste de las proteínas reclutadas para interactuar con distintos compuestos químicos' (de Lorenzo y Pérez-Martín, 1996). En el caso de NahR, podríamos estar ante un regulador que ya había refinado su respuesta y por ello cualquier mutación amplia su espectro de reconocimiento de señales. Esta habilidad de los reguladores que controlan rutas de catabolismo de compuestos aromáticos parece estar bastante extendida. Por ejemplo, VanR, un regulador de la familia GntR que se encuentra en *Cauolobacter crescentus*, es capaz de responder a una gran variedad de compuestos aunque el vanilato es ante el que mejor responde (Thanbichler *et al.*, 2007). YodB, un regulador de la familia MarR que se encuentra en *Bacillus subtilis* puede ser inducido por catecol, 2-metil-hidroquinona y cromanon (Leelakriangsak *et al.*, 2008).

En algunos sistemas el compuesto intermedio de la ruta catabólica expresada ha sido seleccionado como el mejor inductor del regulador. Este es el caso de BenM en *Acinetobater baylyi* ADP1, que es inducido por *cis,cis*-muconate (Ezezika *et al.*, 2007). Se ha sugerido, por estudios teóricos, que este podría ser el mejor compromiso para el circuito regulatorio en términos de robustez, estabilidad y capacidad de respuesta. Esto es debido a que, por un lado, cuando el inductor es un compuesto intermedio de la ruta catabólica aumenta la estabilidad del sistema, y por otro, la mayor robustez y capacidad de respuesta se consiguen cuando el inductor es el primer sustrato. Si el inductor es el producto final, el circuito tiene la menor capacidad de respuesta y robustez posible (Wall *et al.*, 2004).

**Co-evolución entre operones catabólicos y mecanismos reguladores**

Rutas catabólicas diferentes pueden usar mecanismos de regulación similares (como ocurre con el plásmido TOL y el operón *dmp* del plásmido pVI150) o distintos (como el plásmido OCT) para hacer el mismo trabajo. Por otro lado, la misma ruta en diferentes especies también puede estar controlada por distintos mecanismos. En *Alcanivorax Borkumensis*, una bacteria especializada en la asimilación de hidrocarburos alifáticos, existen dos genes que codifican alcano-hidroxilasas: *alkB1* y *alkB2* (van Beilen *et al.*, 2004). Ambos genes son inducidos por alcanos pero la expresión de *alkB1* es mucho mayor. Delante de los promotores de ambos genes existe una repetición invertida similar al sitio de unión del regulador AlkS de *Pseudomonas putida* GPo1. Pero, al contrario de lo que ocurre en *Pseudomonas putida* GPo1, la expresión de *alkS* en *Alcanivorax borkumensis* no está inducida por alcanos y el sitio de unión de AlkS no está presente antes del extremo 5' del promotor de *alkS*. Esto indica que los genes de *A. borkumensis* están regulados utilizando una estrategia molecular distinta de la utilizada en *P. putida* GPo1, pero el fenotipo es el mismo. A este fenómeno se le ha llamado 'gato negro/gato blanco' en referencia a que diferentes mecanismos producen el mismo resultado y no es un caso aislado en las regulación de las rutas catabólicas de degradación de compuestos aromáticos (Cases y de Lorenzo, 2001). El proceso que se propone de evolución de las rutas catabólicas se desarrolla en tres planos que definen su margen de actividad en el medio ambiente. Primero, el ensamblaje de un conjunto de enzimas catabólicas capaces de generar una cantidad de energía suficiente metabolizando un compuesto. Segundo, a partir de un promotores constitutivos, o semi-constitutivos, se adquiere especificidad por la respuesta residual de una pareja regulador-promotor poco específica que sea capaz de

responder a ese compuesto. La elección de una pareja de determinada podría depender más de la historia evolutiva de los operones catabólicos que de las propiedades específicas de las proteínas reguladoras; de hecho es frecuente encontrar operones implicados en biodegradación muy similares controlados por diferentes tipos de promotores (van der Meer *et al.*, 2002). Tercero, en el medio ambiente altamente competitivo que se encuentra en lugares contaminados, los promotores deben ser capaces de procesar diferentes señales ambientales para evitar que la expresión de las rutas catabólica que degrada un compuesto presente en el medio no esté en contra de la adaptación ecológica de la célula al medio. Con este propósito, es esencial que la maquinaria de transcripción de los promotores específicos sea capaz de detectar el estado fisiológico de la célula como conjunto y reaccionar en consecuencia. El refinamiento de la regulación específica y el acoplamiento a la fisiología de la célula son dos procesos que parecen solaparse en el tiempo.

En esta tesis vamos a estudiar los circuitos de regulación y sus componentes de una forma sistemática tomando *Escherichia coli* como modelo para su comparación. De esta forma, pretendemos profundizar en la comprensión del funcionamiento de los circuitos y los factores que determinan su composición y diseño. También, nos interesaremos en la forma en que la conectividad afecta a la organización en operones de los genes catabólicos. Por otro lado, nos fijaremos en los factores sigma por su papel en la integración de los promotores catabólicos con la fisiología del huésped. A continuación, observaremos la organización genética de los genes catabólicos relacionándola con la movilidad de estos genes entre especies. Finalmente, estudiaremos la especificidad de los reguladores con respecto a sus rutas catabólicas reguladas para intentar profundizar en la comprensión de lo que se ha llamado 'ruido regulatorio' y de la co-evolución entre operones catabólicos y mecanismos reguladores.

# OBJETIVOS

• Recopilar toda la información disponible sobre mecanismos de regulación de la transcripción en rutas catabólicas de compuestos aromáticos, integrarla con la información disponible sobre el metabolismo y crear una base de datos para almacenarla

• Hacer la base de datos accesible a la comunidad científica a través de un servidor web

• Caracterizar los circuitos regulatorios que determinan la expresión de los operones catabólicos y sus componentes

• Estudiar la organización genética de los operones catabólicos y las fuerzas que la han modelado

• Estudiar como se integran el metabolismo y la regulación transcripcional en un contexto evolutivo y funcional.

# MÉTODOS

## Extracción de las secuencias de proteínas y genes y el contexto genómico de los organismos

Tomamos como punto de partida para la obtención de la información sobre las secuencias de los factores de transcripción y sus operones regulados dos revisiones sobre la regulación transcripcional de bacterias implicadas en procesos de biodegradación (Tropel y Van der Meer, 2004; Díaz y Prieto, 2000). De aquí obtuvimos los números de acceso a bases de datos del NCBI (Wheeler *et al.*,2007). En la mayoría de los casos eran referencias a GenPept, que es la base de datos que almacena secuencias de aminoácidos y, por lo tanto, proteínas. Partiendo de aquí y utilizando programas en Perl (The Source of Perl, 2007) y módulos de Bioperl (Stajich *et al.*,2002) obtuvimos las secuencias de las proteínas reguladoras, las de sus genes y las de su contexto genómico (el genoma completo en el caso de que estuviera disponible). En algunos casos la referencia a la base de datos era al gen. En esos casos partiendo del gen obtuvimos la proteína. En todos los casos extrajimos el nombre e identificador de Taxonomy (la base de datos de taxonomía del NCBI) de la entrada de GenPept (o GenBank en su defecto).

## Identificación de las secuencias de los sitios de unión al ADN de los reguladores y de los lugares de iniciación de la transcripción de los operones

La revisión de Tropel *et al.* (Tropel y Van der Meer, 2004) contenía, en algunos casos, información sobre los sitios de unión al ADN en forma de secuencias consenso con coordenadas con respecto al lugar de inicio de la transcripción del operón regulado. Para localizar con mayor fiabilidad estos sitios de unión, y también para encontrar los que no figuraban en esta revisión ni en la de Díaz y Prieto (Díaz y Prieto, 2000) , nos dirigimos a los artículos donde se había descrito originalmente la interacción. Para localizarlos nos valimos de dos métodos: la referencia a Pubmed (Wheeler *et al.*,2007) que aparece en las

entradas a la base de datos a las que se hace referencia en las revisiones y, en el caso de no localizar la información de esta manera, búsquedas manuales en Pubmed por el nombre del organismo, el regulador, los genes del operón o cualquier combinación de ellos. A continuación localizamos la secuencia descrita en el artículo dentro de la entrada de GenBank correspondiente por medio de un programa de Perl y, en el caso de no localizarla, comprobamos manualmente que no era un error de anotación. De manera similar buscamos en los artículos los inicios de transcripción de los operones. En algunos ocasiones los inicios de transcripción también venían descritos en la entrada de GenBank. En estas situaciones, si no coincidían las dos fuentes de información, aceptábamos como correcta la información descrita en el artículo.

## Obtención de la información sobre los efectores de los reguladores

Al igual que en el caso de los inicios de transcripción extrajimos la información sobre los efectores de las revisiones y posteriormente la contrastamos y ampliamos con los artículos donde se describieron originalmente.

## Integración con la base de datos de metabolismo

Antes de que comenzáramos con la construcción de la base de datos sobre regulación ya se había extraído gran cantidad de información sobre el metabolismo de la "red global de biodegradación" por lo que decidimos integrar ambas fuentes de información.
Por un lado, los datos a integrar sobre metabolismo asignaban a cada reacción descrita en la base de datos sobre biodegradación de la Universidad de Minnesota (Ellis *et al.*, 2006) uno o varios complejos enzimáticos. Estos complejos enzimáticos estaban formados por proteínas cuya secuencia e información asociada se extraía de GenPept. Por el otro, algunos genes de los operones que obtuvimos buscando información sobre regulación producen proteínas que forman parte de complejos enzimáticos. Estos complejos enzimáticos coincidían con frecuencia con los que estaban ya asignados en la base de datos sobre metabolismo por lo que debíamos unificar las dos fuentes. Los pasos dados y los métodos que utilizamos para ello fueron los siguientes:

• En primer lugar comparamos los números de acceso a GenPept de las proteínas eliminando las que tenían el mismo para quedarnos sólo con una copia de cada una.
• A continuación localizamos las entradas idénticas en ambas fuentes por medio del programa MagicMatch (Smith *et al.*,2005) que genera "huellas digitales" a partir de la secuencia de aminoácidos lo que permite identificar las proteínas que son iguales.
• Posteriormente comprobamos si las proteínas idénticas pertenecían al mismo organismo consultando las entradas de GenPept manualmente y si, por lo tanto, son la misma proteína aunque tengan un identificador distinto. En el caso de que fueran la misma proteína nos quedábamos con la copia con el registro de entrada más moderno.
•Tras esto realizamos una comparación de las secuencias con BLAST (Altschul *et al.*, 2005) , un algoritmo que permite comparar secuencias de aminoácidos o de nucleótidos, para hacer una inspección definitiva de las proteínas y localizar las que, aún siendo la misma proteína del mismo organismo, no hubiéramos localizado por ser la secuencias ligeramente diferentes entre ellas por algún error al anotarlas.

## Asignación de proteínas a complejos enzimáticos y estos a su vez a reacciones

Tras acabar el proceso de integración quedaron proteínas que no estaban todavía asignadas al complejo enzimático al que pertenecen. Para facilitar la tarea de asignarlas utilizamos dos herramientas de 'clustering': BLATCLUST (Altschul *et al.*, 2005) y TRIBE-MCL (Enright *et al.*, 2002) . Estas aplicaciones forman grupos de proteínas similares. Al utilizarlas con toda nuestra lista de proteínas agrupamos proteínas que no tienen complejo enzimático asociado con otras que sí lo tienen. En numerosos casos los complejos enzimáticos de las proteínas que forman parte de un mismo grupo son similares lo que nos permite agilizar el proceso de definición de los complejos. Para asegurarnos de que los complejos estaban definidos correctamente contrastamos la información anotada por nosotros con la que aparece en las entradas de GenPept y , finalmente, con los artículos de donde extrajimos los datos originalmente.
De manera similar, los grupos formados mediante el 'clustering' nos ayudan a asignar más rápidamente los complejos enzimáticos a las reacciones que catalizan. En este caso también comprobamos por medio de las anotaciones de GenPept y los artículos originales que nuestra información es correcta.

# Modelo molecular

Para poder almacenar y gestionar adecuadamente toda la información surgida de la integración de los datos sobre metabolismo y regulación decidimos construir una base de datos. Para ello tuvimos primero que realizar un modelo molecular que describiera adecuadamente la información que queríamos almacenar y que está descrito en la figura 1.

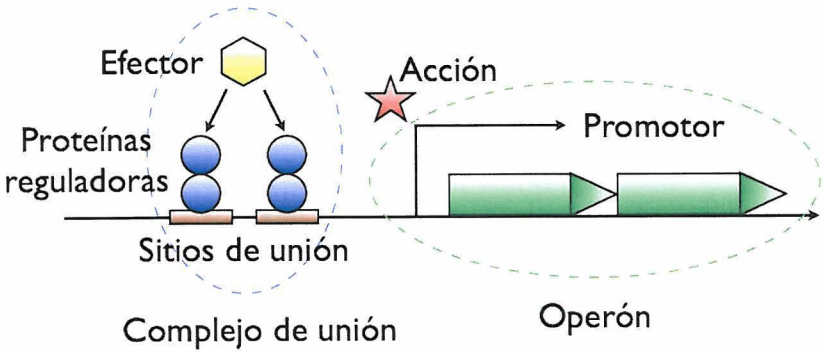



**Figura 1.** Modelo molecular de un sistema de regulación de la transcripción. Los complejos de unión ejercen una acción sobre el promotor alterando la expresión de los genes contenidos en el operón

Nuestra intención fue crear un modelo flexible, que nos permitiera almacenar todos los detalles disponibles, pero robusto para poder trabajar y realizar cálculos con poca información.

Uno de los aspectos más destacables del modelo es su capacidad para contener información molecular muy detallada ya que describe por separado todas las partes que intervienen en el proceso que conduce a la expresión, o inhibición si ese es el caso, de unos genes regulados: proteínas reguladoras, efectores, sitios de unión al ADN, promotor, operón (que en nuestro modelo amplia la definición tradicional e incluye como operones todos los genes que se transcriben solos, como es el caso de muchos de los reguladores) y los propios genes regulados. Esta separación en componentes nos permite llegar al detalle de especificar la estequiometría de cada elemento pero con la flexibilidad de poder incluir la información disponible aunque esta no sea completa.

# Construcción de la base de datos

A continuación convertimos el modelo en una base de datos utilizando para ello PostgreSQL (PostgreSQL, 2007), un sistema de bases de datos relacionales. El diseño aparece detallado en un gráfico en la siguiente página. Las unidades que componen la base de datos relacional son tablas que contienen datos y que están conectadas entre sí por relaciones. Vamos a describir únicamente las tablas que contienen información relacionada con la regulación de la transcripción, aunque la base de datos también contiene tablas que contienen información sobre el metabolismo.

Para ilustrar el proceso de construcción de la base de datos comenzaremos describiendo la tabla "protein" e iremos enlazando con las demás tablas según su relación con las que vayan siendo descritas con anterioridad. Las relaciones entre tablas se pueden ir siguiendo en el gráfico de la siguiente página. En este gráfico también aparecen detallados todos los campos de las tablas que, en algunos casos, únicamente se mencionan.

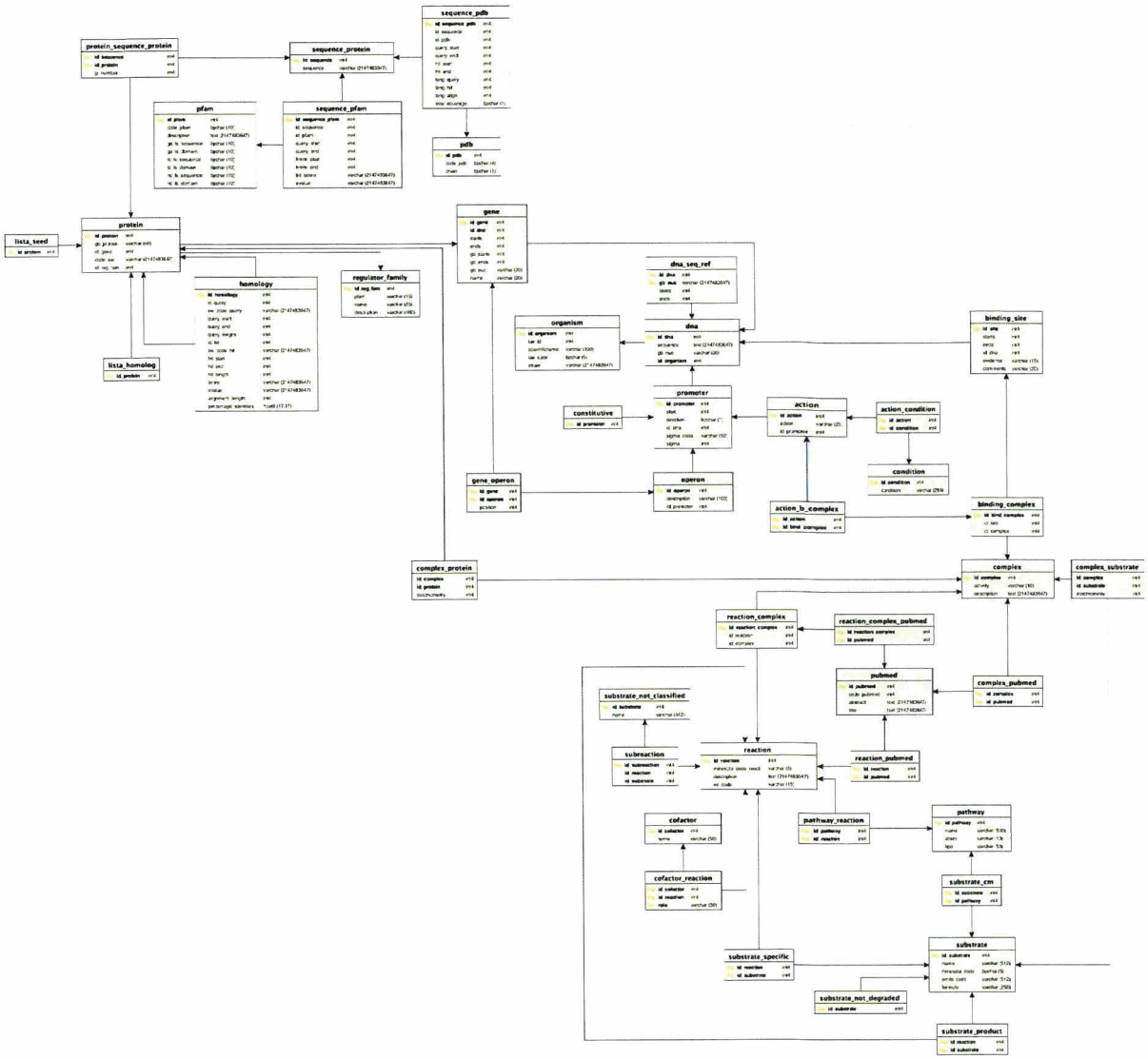



**Figura 2.** Diagrama de la base de datos Bionemo. En el diagrama se pueden observar todos los campos contenidos en las tablas y las conexiones entre ellas. Cuando una tabla tiene un campo que hace referencia a otra esta conexión está representada por una flecha que va de la primera a las segunda

**Proteínas, familias de reguladores y secuencias de aminoácidos.**

| Tabla 'protein' | | | |
|---|---|---|---|
| **Campo** | **Tipo** | **Descripción** | **Ejemplo** |
| id_protein | Número entero | Identificador | 1 |
| gb_protein | Texto variable | Identificador GenPept | BAC92712 |
| id_gene | Número entero | Identificador | 1 |
| code_sw | Texto variable | Identificador Swissprot o Tremble | trl:Q75WN5_RHOSR |
| id_reg_fam | Número entero | Identificador | 1 |

En la tabla "protein" el primer campo, "id_protein", es el identificador de la entrada en nuestra base de datos, "gb_protein" es el identificador de la entrada para la proteína en GenPept (Wheeler *et al.*, 2007) , "id_gene" relaciona la entrada de la proteína con la entrada del gen que la expresa en la tabla "gene" y "code_sw" es el identificador de Swissprot (Boeckmann *et al.*, 2003) ,una base de datos con información sobre proteínas que está revisada manualmente. Introdujimos el identificador de Tremble (Boeckmann *et al.*, 2003) en este campo en el caso de no disponer de un identificador de Swissprot. Tremble es una base de datos de proteínas obtenidas al traducir secuencias de genes a aminoácidos y que no está revisada manualmente. El campo "id_reg_fam" relaciona la tabla "protein" con la tabla "regulator_family", que contiene la información sobre las familias de reguladores. Es necesario aclarar que este campo permanece vacío en todos los casos en los que la proteína introducida no es un regulador.

Las secuencias de las proteínas se almacenan por separado en la tabla "sequence_protein". Hacemos esto para evitar la redundancia de la información, ya que dos proteínas ortólogas que pertenezcan a diferentes organismos pueden tener la misma secuencia. De este modo tenemos dos entradas en la tabla "protein" para las dos

proteínas distintas relacionadas con una única entrada en la tabla "sequence_protein" a través de una tabla intermedia.

| Tabla 'regulator_family' | | | |
|---|---|---|---|
| **Campo** | **Tipo** | **Descripción** | **Ejemplo** |
| id_reg_fam | Número entero | Identificador | 1 |
| pfam | Texto variable | Identificador Pfam | PF00072 |
| name | Texto variable | nombre | RR_TCST |
| description | Texto variable | Descripción de la familia de reguladores | Response Regulator. Two-component regulatory systems |

En la tabla "regulator_family" está contenida la información sobre las familias regulatorias a las que pertenecen los reguladores de nuestra base de datos. Esta clasificación por familias es la que aparece en las revisiones de donde obtuvimos la información originalmente (Tropel y Van der Meer, 2004; Díaz y Prieto, 2000). Junto con el nombre, en esta tabla se almacena el identificador de Pfam (Bateman *et al.*,2004) para la familia de proteínas de Pfam a la que pertenece la familia de reguladores.

**Genes y secuencias de ADN**

| Tabla 'gene' | | | |
|---|---|---|---|
| **Campo** | **Tipo** | **Descripción** | **Ejemplo** |
| id_gene | Número entero | Identificador | 1 |
| id_dna | Número entero | Identificador | 1 |
| starts | Número entero | Coordenada | 1305 |
| ends | Número entero | Coordenada | 2675 |
| gb_starts | Número entero | Coordenada | 1305 |
| gb_ends | Número entero | Coordenada | 2675 |
| gb_nuc | Texto variable | Identificador Genbank | AB120955.1 |
| name | Texto variable | Nombre del gen | etbA1 |

De la tabla "gene" cabe destacar la presencia de un identificador que asocia el gen a una entrada de GenBank (Wheeler *et al.*, 2007) , "gb_nuc", y también la existencia de dos tipos de coordenadas. Las coordenadas "gb_starts" y "gb_ends" relacionan el gen con sus coordenadas en la entrada de GenBank introducida en "gb_nuc". Las coordenadas "starts" y "ends" relacionan el gene con la entrada de la tabla "dna" introducida en "id_dna". Introdujimos estos dos tipos de coordenadas porque en algunos casos la secuencia de ADN de la tabla "dna" no se corresponde con ninguna entrada de GenBank porque ha sido editada por nosotros para añadir información obtenida a través de artículos pero, a pesar de ello, queremos conservar las coordenadas con respecto a GenBank. La hebra de ADN que codifica el gen viene definida por los valores de "starts" y "ends": si "starts" es menor que "ends" es la hebra positiva, si no es la negativa.

| Tabla 'dna' | | | |
|---|---|---|---|
| **Campo** | **Tipo** | **Descripción** | **Ejemplo** |
| id_dna | Número entero | Identificador | 30 |
| sequence | Texto variable | Secuencia de ADN | ATGCATTT TTGGG... |
| gb_nuc | Texto variable | Identificador Genbank | AB120955.1 |
| id_organism | Número entero | Identificador | 30 |

La tabla "dna" contiene las secuencias de ADN, como mencionamos anteriormente, y las relaciona con el organismo al que pertenecen a través del campo "id_organism". También aquí encontramos un identificador de GenBank, "gb_nuc", que en este caso puede ser nulo si la secuencia no coincide exactamente con una entrada de GenBank por haber sido editada para incluir información extraída de un artículo, como mencionamos al describir la tabla gen.

**Organismos**

| Tabla 'organism' | | | |
|---|---|---|---|
| **Campo** | **Tipo** | **Descripción** | **Ejemplo** |
| id_organism | Número entero | Identificador | 1 |
| tax_id | Texto variable | Identificador Taxonomy | 101510 |
| scientificname | Texto variable | Nombre científico | Rhodococcus sp. |
| tax_code | Texto variable | Identificador BLAST | RHOSR |

| Tabla 'organism' | | | |
|---|---|---|---|
| strain | Texto variable | Nombre de la cepa | RHA1 |

En la tabla "organism" se almacena el nombre científico de la especie, "scientificname", la cepa, "strain", y dos tipos de identificadores de otras bases de datos. Para asociar nuestra entrada con la correspondiente de la base de datos Taxonomy del NCBI (Wheeler *et al.*, 2007) utilizamos el campo "tax_id" y el campo "tax_code" incluye el identificador utilizado en Swissprot (Boeckmann *et al.*, 2003) para diferenciar entre especies.

**Sustratos y efectores**

Los sustratos, tanto los que sirven como inductores como los que utilizan las reacciones, se almacenan en la tabla "substrate". Los campos a destacar en esta tabla son el identificador de la base de datos de biodegradación de Minnesota (Ellis *et al.*, 2006) , "minesota_code", el código SMILES de la molécula (un código que permite especificar la estructura de una molécula con una cadena corta de caracteres ) y la fórmula.

| Tabla 'sustrate' | | | |
|---|---|---|---|
| **Campo** | **Tipo** | **Descripción** | **Ejemplo** |
| id_substrate | Número entero | Identificador | 1 |
| name | Texto variable | Nombre del compuesto | Acetamide |
| minesota_code | Texto variable | Identificador de UM-BBD | c0658 |
| smile_code | Texto variable | Fórmula con estructura | CC(=O)N |
| formule | Texto variable | Fórmula | C<sub>2</sub>H<sub>5</sub>NO |

### Complejos, sitios de unión al ADN y complejos de unión

Para explicar la estructura de las tablas que definen los complejos retomamos la descripción del modelo molecular. En la figura 3 se representa un complejo.

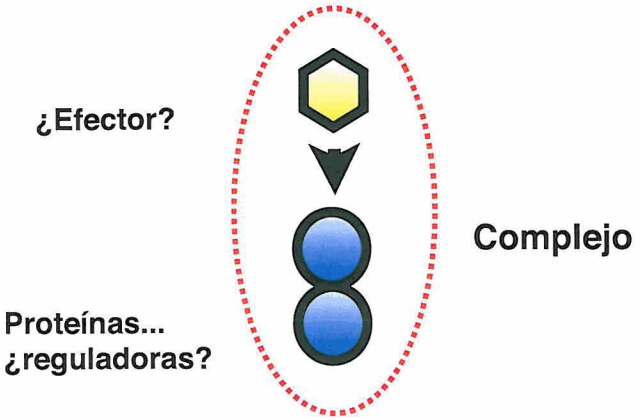



**Figura 3. Modelo molecular de un complejo.** Definimos complejo como una asociación de proteínas y, en algunos casos, una molécula efectora. Existen dos tipos de complejos: reguladores y enzimáticos. Los enzimáticos están compuestos únicamente de proteínas (los cofactores no se consideran parte del complejo) y los reguladores pueden incluir una molécula inductora de la acción que realizan o no, dependiendo de que ese sea el caso concreto o no.

Los complejos pueden ser enzimáticos o reguladores. Los complejos enzimáticos están compuestos solamente de proteínas. Los complejos reguladores pueden estar formados solo de proteínas o incluir también una molécula efectora de la acción que realizan. La actividad que realizan los complejos se especifica en la tabla complex en el campo "activity": si es regulador el valor es "REG" y si es enzimático "ENZ".

| Tabla 'complex' | | | |
|---|---|---|---|
| **Campo** | **Tipo** | **Descripción** | **Ejemplo** |
| id_complex | Número entero | Identificador | 2 |
| activity | Texto variable | actividad | REG |
| description | Texto | Descripción del complejo | CatR dimer + cis,cis-muconate |

Para relacionar los complejos con las subunidades que los forman existen dos tablas intermedias: "complex_protein", en la que se especifican las proteínas que forman parte del complejo, y "complex_substrate", en la que se indica el inductor. En ambas tablas

existe un campo en el que se puede especificar cuantas subunidades de ese tipo forman parte del complejo, "stoichiometry". En el ejemplo de la figura 3 la estequiometría sería 2 para las proteínas y 1 para los sustratos.

La información sobre los sitios de unión al ADN se almacena en la tabla binding_site (siguiente página). La hebra de ADN que se asigna a los sitios de unión es la que está anotada en el archivo GenBank del que se extrajo la información. Si un sitio de unión interviene en la regulación de un promotor que se transcribe en la dirección opuesta se asigna en la dirección en que viene descrito en el archivo de GenBank, aunque asumimos que el sitio de unión forma parte de la hebra complementaria.

En la tabla "binding_site" las coordenadas con respecto a la secuencia de ADN, que se especifica en "id_dna", se indican en los campos "starts" y "ends". En "evidence" se describe si la evidencia anotación es putativa o experimental y en "comments" se añaden comentarios (RBS, IHF, etc...).

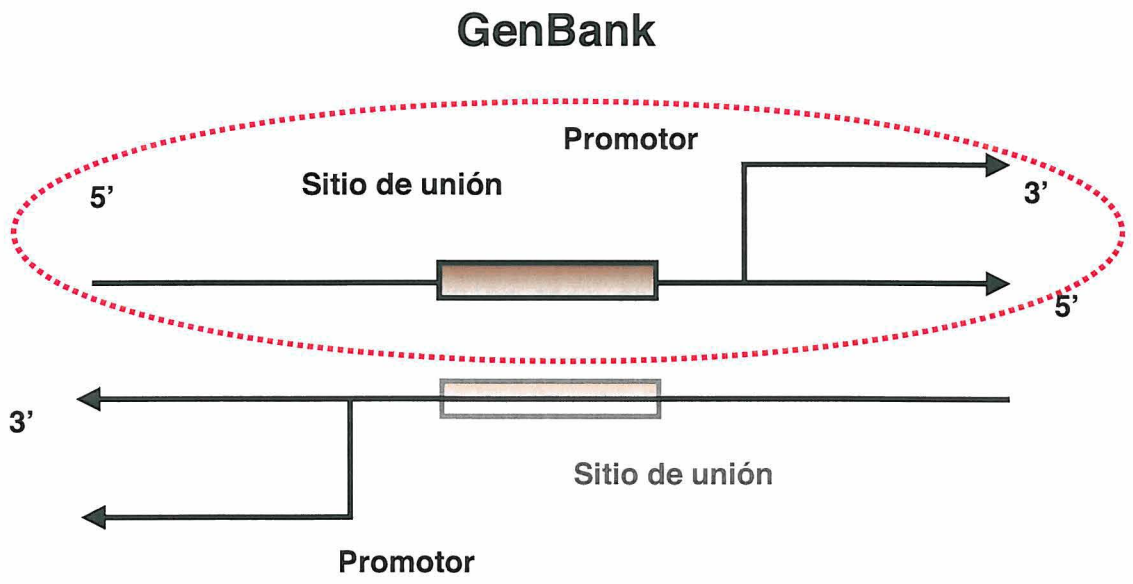



**Figura 4. Sitios de unión, hebra de ADN y archivos GenBank.** Los sitios de unión de los complejos están asignados a la hebra de ADN que está anotada en el archivo de GenBank del que extrajimos la información. Puede darse el caso de que un sitio de unión regule un promotor que se encuentre en la otra hebra. Aunque nosotros lo tengamos anotado en la hebra complementaria asumimos que ese sitio de unión se encuentra también en la otra hebra.

| Tabla 'binding_site' | | | |
|---|---|---|---|
| **Campo** | **Tipo** | **Descripción** | **Ejemplo** |
| id_site | Número entero | Identificador | 2 |
| starts | Número entero | Coordenada | 1305 |
| ends | Número entero | Coordenada | 2675 |
| id_dna | Número entero | Identificador | 30 |
| evidence | Texto variable | evidencia | Putative |
| comments | Texto variable | comentarios | RBS |

En la figura 5 se ilustra nuestra definición de complejo de unión. En resumen, un complejo de unión sería uno de los complejos regulatorios que hemos introducido en la tabla "complex" unido a uno de los sitios de unión al ADN que hemos introducido en la tabla "binding_site". La asociación se hace a través de la tabla "binding_complex" que contiene un identificador del complejo, "id_complex", y uno del sitio de unión, "id_site".

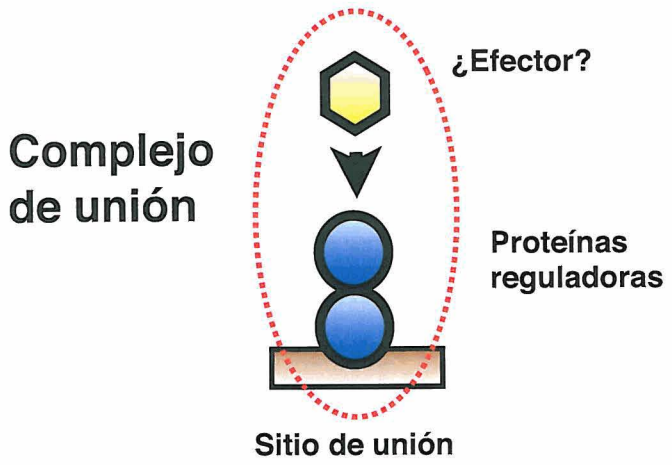



**Figura 5. Modelo molecular de un complejo de unión.** Un complejo de unión está formado por un sitio de unión al ADN con el que está asociado un grupo de proteínas reguladoras. Este grupo de proteínas reguladoras puede estar relacionado con una molécula inductora en algunos casos aunque no siempre.

## Promotores y operones

La definición de promotor que utilizamos en nuestra base de datos está ilustrada en la figura 6. Un promotor es el inicio de transcripción de una ARNpolimerasa asociado a un factor sigma que recluta la ARNpolimerasa. El mismo inicio de transcripción asociado a un factor sigma diferente se considera un promotor distinto.

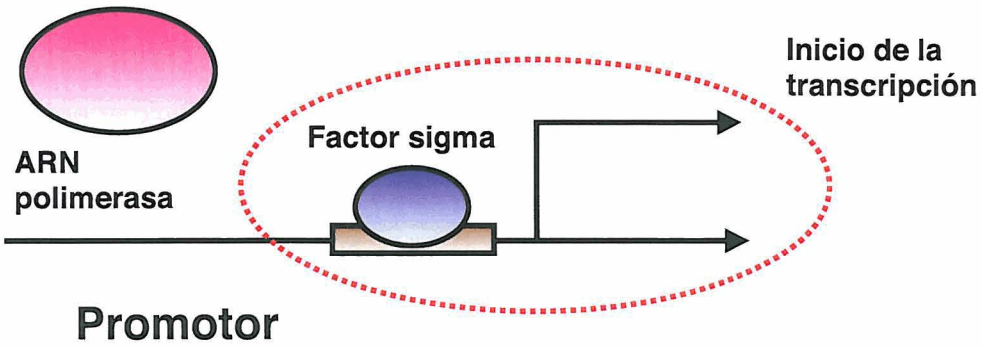



**Figura 6. Modelo molecular de un promotor.** Un factor de transcripción sigma reconoce una secuencia por la que tiene afinidad. La ARNpolimerasa reconoce el factor sigma y se une a él iniciando la transcripción tras liberarlo. En nuestro modelo un promotor se define por el punto donde se inicia la transcripción y el factor sigma que recluta la ARNpolimerasa.

| Tabla 'promoter' | | | |
|---|---|---|---|
| **Campo** | **Tipo** | **Descripción** | **Ejemplo** |
| id_promoter | Número entero | Identificador | 2 |
| start | Número entero | Coordenada | 1305 |
| direction | Texto | Dirección de la transcripción | + |
| id_dna | Número entero | Identificador | 30 |
| sigma_class | Número entero | Tipo de sigma | 70 |
| sigma | Número entero | Identificador de protein | 2082 |

En la tabla 'promoter' se indican el inicio de la transcripción en el campo "start" y la dirección de la transcripción en "direction". La secuencia de ADN en la que se encuentra el promotor se encuentra en la tabla "dna" y su identificador está en el campo "id_dna". En el campo "sigma_class" se especifica la familia de sigma a la que pertenece el factor sigma asociado a este promotor. En "sigma" se indica el "id_protein" para el factor sigma, que es una proteína y por lo tanto está almacenado en la tabla "protein". En la tabla "constitutive" , que está relacionada con la tabla "promoter", hay una lista de identificadores de promotores que se expresan constitutivamente.

La figura 7 ilustra nuestro modelo de operón. Como ya comentamos anteriormente, nuestra definición de operón abarca los genes que se transcriben en solitario, como es el caso de muchos de los reguladores. El mismo grupo de genes constituirían dos operones diferentes si se expresan desde promotores diferentes. Asimismo, dos operones distintos pueden contener el mismo promotor que expresa un grupo diferente de genes.

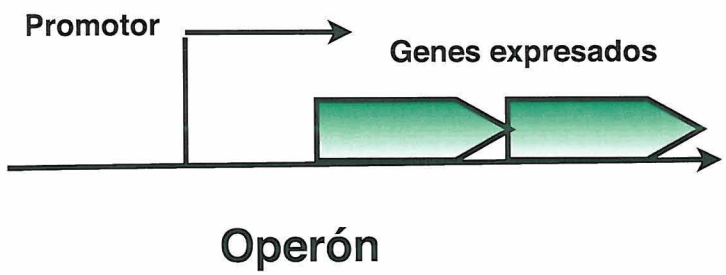



**Figura 7. Modelo molecular de un operón.** Definimos operón como un grupo de genes expresados desde un mismo promotor y que tienen un orden definido por la dirección de la transcripción. Si no se conoce el promotor el operón queda definido por el grupo de genes.

| Tabla 'operon' | | | |
|---|---|---|---|
| **Campo** | **Tipo** | **Descripción** | **Ejemplo** |
| id_operon | Número entero | Identificador | 216 |
| description | Texto variable | Descripción del operón | areCBA |
| id_promoter | Número entero | Identificador | 2 |

En la tabla "operon" el campo "id_promoter" relaciona el operon con la tabla "promoter". En el caso de que no haya un promotor descrito para el operón este campo puede quedar

vacío. La asignación de los genes se hace a través de una tabla intermedia, "gene_operon", que conecta la entrada del gen en la tabla "gene" con el operón. La posición del gen en el operón está indicada en "position" y está relacionada con el orden en que se transcriben los genes como se puede ver en la figura 7. El mismo gen puede tener posiciones distintas en distintos operones si estos están expresados desde diferentes promotores.

## Acciones y condiciones

El modelo molecular que explica las acciones está relacionado con el ejemplo de la figura 2 que describe una activación. Los elementos que definen una acción vienen descritos en la figura 8, que resulta útil para integrar la explicación de los componentes de la acción por separado y el modelo de la base de datos relacional que almacena la información.

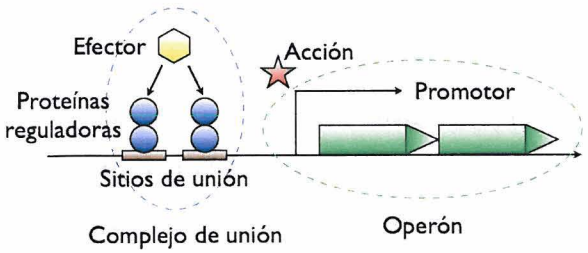
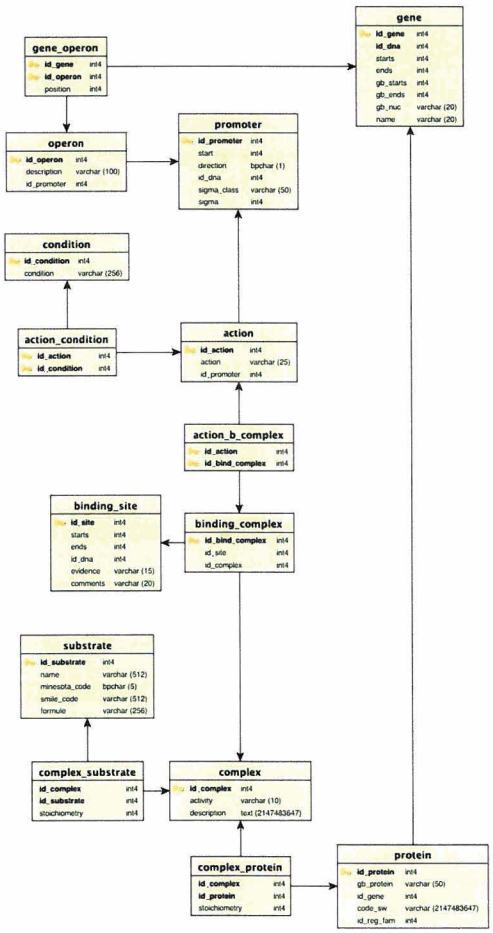



**Figura 8. Modelo molecular de un acción y su codificación en la base de datos.** Una acción es el cambio en la expresión de un promotor causado por la presencia de uno o varios complejos de unión Cada complejo de unión está formado por un sitio de unión al ADN una o varias proteínas reguladoras y, opcionalmente, una molécula inductora . La acción hace que se active o reprima la expresión de un promotor que transcribe para uno o varios operones. Los operones contienen los genes sobre los que recae la acción regulatoria.



| Tabla 'action' | | | |
|---|---|---|---|
| **Campo** | **Tipo** | **Descripción** | **Ejemplo** |
| id_action | Número entero | Identificador | 1 |
| action | Texto variable | Descripción de la acción | activation |
| id_promoter | Número entero | Identificador | 77 |

En la tabla "action" anotamos el promotor sobre el que recae la acción en "id_promoter", que relaciona la tabla con la tabla "promoter". Para asociar los complejos de unión a la acción en la que intervienen está la tabla intermedia "action_b_complex" que relaciona las tablas "action" y "binding_complex".

| Tabla 'condition' | | | |
|---|---|---|---|
| **Campo** | **Tipo** | **Descripción** | **Ejemplo** |
| id_condition | Número entero | Identificador | 1 |
| condition | Texto variable | Descripción de la condición | synergistic activation |

En la tabla "condition" se almacenan otras condiciones que se tienen que dar para que se produzca la acción o características de la acción que son interesantes y no se almacenan en ninguna otra tabla. Existe una tabla intermedia, "action_condition", para conectar las acciones con las condiciones ya que una misma condición puede ser requerida por varias acciones o una misma acción puede requerir varias condiciones.

## Homología

Para poder realizar posteriormente experimentos en los que comparamos proteínas entre sí, realizamos un BLAST (Altschul *et al.*, 2005) entre las proteínas almacenadas en nuestra base de datos frente a las bases de datos Swissprot y Trembl en su versión de enero del 2007. Posteriormente incorporamos a nuestra base de datos local con nuestras proteínas y las que tenían un 'e-value' menor de $1*10^{-6}$. De nuevo realizamos un BLAST entre todas ellas y almacenamos los resultados en la tabla 'homology'. El identificador de la proteína que se compara con el resto de la base de datos se guarda en el campo 'id_query' y el de la proteína con la que se compara se guarda en 'id_hit'. El código Swissprot o Trembl de cada una se guarda en 'sw_code_query' y 'sw_code_hit' respectivamente. La longitud de cada una de ellas se guarda en 'query_length' y 'hit_length'. La longitud en número de aminoácidos de la zona alienada entre las dos proteínas se guarda en el campo 'alignement_length' y las coordenadas de la zona alineada son 'query start' y 'query_end' para la primera proteína y 'hit start' y 'hit_end' para la segunda. Los valores de identidad se guardan en los campos 'score', 'evalue' y 'percentage_identities'. Los campos 'score' y 'evalue' son valores de puntuación del programa BLAST que describen la similitud entre las secuencias según unos parámetros determinados. A mayor 'score' ó menor 'evalue' mayor identidad entre secuencias. En 'percentage_identities' se guarda el porcentaje de aminoácidos idénticos entre las dos proteínas dentro de la zona alineada.

| Tabla 'homology' | | | |
|---|---|---|---|
| **Campo** | **Tipo** | **Descripción** | **Ejemplo** |
| id_homology | Número entero | Identificador | 40644 |
| id_query | Número entero | Identificador | 828 |
| sw_code_query | Texto variable | Identificador Swissprot o Tremble | swl:BENC_ACIAD |
| query_start | Número entero | Coordenada | 14 |

| Tabla 'homology' | | | |
|---|---|---|---|
| query_end | Número entero | Coordenada | 348 |
| query_length | Número entero | Longitud de la proteína | 348 |
| id_hit | Número entero | Identificador | 1333 |
| sw_code_hit | Texto variable | Identificador Swissprot o Tremble | trl:Q0WZ45_ PSEPU |
| hit_start | Número entero | Coordenada | 3 |
| hit_end | Número entero | Coordenada | 335 |
| hit_length | Número entero | Longitud de la proteína | 336 |
| score | Texto | BLAST-score | 381 |
| evalue | Texto | BLAST e-value | 2,00E-107 |
| alignement_length | Número entero | Longitud del alineamiento | 335 |
| percentage_identities | Número decimal | porcentaje de aa alineados | 55,22 |

**Artículos**

Los identificadores de Pubmed (Wheeler *et al.*,2007) de los artículos de los que se extrajo la información se encuentran almacenados en la tabla "pubmed". La tabla intermedia "complex_pubmed" conecta cada complejo regulatorio con los artículos de los que se sacó la información para definirlo.

**Construcción de un interfaz para la base de datos**

Debido a la complejidad del diseño de la base de datos, tareas sencillas se convierten en difíciles y tediosas. Por ejemplo, para consultar a que organismo pertenece la proteína reguladora que tiene el identificador 1 en la tabla "protein" tenemos que ejecutar la siguiente consulta:

```
SELECT s.scientific_name FROM protein p, gene g, dna d, organism o,
WHERE p.id_gene = g.id_gene AND g.id_dna = d.id_dna
AND d.id_organism = o.id_organism AND p.id_protein = 1;
```

Alguna consulta algo más elaborada como comprobar en que organismos está representada una ruta metabólica podría ocupar prácticamente una página. Por eso se hace necesario crear alguna herramienta que nos permita interactuar con la base de datos efectuando preguntas complejas de forma sencilla. Esta herramienta es un API (de "Application Program Interface" en inglés). La consulta anterior utilizando la API se escribiría así:

```
get_protein('id',1)->its_organism->display_description
```

El acceso a la base de datos de este modo resulta mucho más sencillo e intuitivo, no precisa de un conocimiento exhaustivo de las tablas implicadas en la consulta ni de sus campos y, además , permite realizar programas en los que consultas complejas se van enlazando de forma sencilla para profundizar en el análisis de los datos. Toda la caracterización de los circuitos de regulación que se describe en la sección más de resultados ha sido realizada valiéndonos de esta herramienta. También nos ha facilitado la tarea de construir un servidor Web para hacer accesible a través de Internet toda la información almacenada.

## Creación de un servidor Web

El desarrollo del mencionado servidor Web fue realizado junto a Almudena Trigo que se encargó de completar las partes directamente relacionadas con el metabolismo. Para implementarlo utilizamos un interfaz de entrada común (CGI, de sus siglas en inglés). Un interfaz de entrada común es una tecnología de la World Wide Web que permite a un cliente (explorador web) solicitar datos de un programa ejecutado en un servidor web. CGI especifica un estándar para transferir datos entre el cliente y el programa. Las aplicaciones que se ejecutan en el servidor reciben el nombre de CGIs. Nuestro servidor web llama a un módulo de aplicación que ejecuta y maneja los modos de ejecución de la aplicación. Cada modo de ejecución se corresponde con una página web. Cada uno de estos modos de ejecución contiene funciones que se encargan de recopilar la información que se ha de servir en cada página. Una vez que el módulo de ejecución ha recopilado la información ejecuta un módulo de que contiene el código HTML de cada una de las páginas del servidor web de Bionemo. Dependiendo de los parámetros que se le pasen una parte del programa será ejecutada (en concreto, aquella que se corresponda con el código HTML de la página a mostrar). Esta estructura nos permite independizar el código Perl (qué información se muestra) del código HTML (cómo se muestra la información)

## Extracción de información sobre *Escherichia coli*

Para poder hacer las comparar la información almacenada sobre los sistemas de biodegradación con la que está disponible sobre *Escherichia coli* construimos una base de datos análoga a Bionemo. Esta base de datos contiene información sobre el catabolismo de *E. coli*.

La información fue extraída de Ecocyc (Keseler *et al.*, 2008) descargando la base de datos localmente. Tras la descarga obtenemos un conjunto de ficheros que contienen la información de las diferentes entidades (genes, proteínas, unidades transcripcionales, rutas metabólicas, etc.) con diversas etiquetas que especifican características de la entidad y las relacionan con otras.

Para filtrar los operones implicados en procesos de degradación seguimos los siguientes pasos.

Partimos de los fichero 'classes.dat'. En este fichero seleccionamos las rutas metabólicas que pertenecen a la categoría 'Degradation ' y que, por lo tanto, están implicadas en procesos de degradación (campo 'TYPES' ). Con esta información obtenemos una jerarquía de tipos de rutas metabólicas implicadas en distintos procesos de degradación ('Other degradation', 'Degradation', 'Secondary metabolite degradation', etc.)

A continuación, seleccionamos en el archivo 'pathway.dat' los identificadores de las rutas metabólicas de degradación (campo 'UNIQUE-ID').

Como siguiente paso, obtenemos la lista de identificadores de las reacciones de cada ruta en el mismo archivo seleccionando los campos 'REACTION-LIST' en cada ruta.

Con el identificador de la reacción se puede obtener la lista de enzimas que la realizan en el archivo 'enzrxns.dat ' seleccionando los campos 'ENZYME ' de las entradas que contienen la reacción anteriormente obtenida en el campo 'REACTION'

Posteriormente buscamos las proteínas que forman la enzima en el archivo 'proteins.dat'. En este archivo localizamos las entrada que corresponden a la categoría 'Protein-Complexes' (campo 'TYPES' ) y dentro de estas entradas obtenemos las proteínas que forman el complejo en el campo 'COMPONENTS' de cada enzima que localizamos por su entrada ('UNIQUE-ID').

Para encontrar los genes que codifican las proteínas utilizamos este mismo archivo 'proteins.dat'. Buscamos las entradas de cada proteína (campo 'UNIQUE-ID') y localizamos el gen que las codifica en el campo 'GENE'.

A continuación le asignamos al gen las unidades transcripcionales en las que está contenido consultando el archivo 'transunits.dat'. Nos quedamos con los identificadores ('UNIQUE-ID') de las unidades transcripcionales que contienen los genes que obtenido anteriormente.

En este mismo archivo obtenemos los promotores de las unidades transcripcionales que están asignados en los campos 'COMPONENTS'.

Para obtener la regulación de las unidades transcripcionales recurrimos al archivo 'regulation.dat'. En este archivo obtenemos los complejos reguladores (campo 'REGULATOR') y el tipo de acción que ejercen sobre el promotor (campo 'MODE') seleccionando las entradas en las que el promotor regulado ('REGULATED-ENTITY') es uno de los asignados a las unidades transcripcionales que obtuvimos anteriormente.

Finalmente, para obtener las proteínas que forman los complejos, sus genes, sus unidades transcripcionales y su regulación de seguimos los mismos pasos que para obtener la información de los componentes de las enzimas.

## Construcción de la base de datos del catabolismo de *Escherichia coli*

Construimos la base de datos del catabolismo de *Escherichia coli* siguiendo el mismo modelo molecular que en Bionemo excepto con una salvedad: no incluimos información sobre los sitios de unión al ADN ya que no la consideramos necesaria para comparar los aspectos que deseábamos comparar entre los circuitos de regulación. Por ello la principal diferencia es que en lugar de asignar complejos de unión a las acciones les asignamos complejos. La estructura de las tablas y de la base de datos, por lo demás, es totalmente análoga por lo que no consideramos necesario explicar de nuevo todos los detalles moleculares y describir de nuevo las tablas.

## Caracterización de los circuitos de degradación

Para comparar los componentes de los circuitos de biodegradación con los de *Escherichia coli* utilizamos la anteriormente mencionada API y la base de datos del catabolismo de *E. coli.*

Para asegurarnos de que realizamos siempre nuestros cálculos con los mismos datos, creamos versiones locales de Genbank (Wheeler *et al.*, 2007), Swissprot y Trembl (Boeckmann *et al.*, 2003). Para crear la versión local de Genbank utilizamos el módulo de perl Bio::Index::Genbank que crea un índice que permite acceder a la base de datos local a través de Bioperl (Stajich *et al.*, 2002). La versión que utilizamos de Genbank es la 156.0 del 15 de Octubre del 2006.

Para calcular la contigüidad de los reguladores a sus genes regulados contamos los genes situados entre ambos que encontrábamos en la entrada de la secuencia de ADN de Genbank: si no había genes entre el gen regulado más cercano y el gen del regulador los considerábamos contiguos.

Finalmente para realizar los tests estadísticos utilizamos el software R para análisis estadístico y gráfico (Ihaka y Gentleman, 1996).

## Conectividad y unidades transcripcionales e integración con la fisiología

De nuevo utilizamos el API y la base de datos local de *Escherichia coli* para realizar los cálculos y el software R para el análisis estadístico (Ihaka y Gentleman, 1996).

En el apartado de la integración con la fisiología, para determinar la homología entre operones expresados desde promotores asociados a sigma54 seguimos los siguientes pasos:

• Para considerar que dos operones fueran homólogos al menos el 66% de los genes del operón más corto debería de tener un determinado umbral de identidad de secuencia entre las proteínas que expresa y las proteínas que expresa el operón con el que lo comparamos (por ejemplo, si comparamos un operón de 3 genes con una de 4 al menos 2 genes del primero deben de ser homólogos a 2 genes del segundo)

• Para considerar dos genes como homólogos elegimos distintos grados de identidad dependiendo del experimento que se especifican, y justifican, en los resultados, pero en todos los casos las proteínas que codifican no deberían diferir en más de 10 aminoácidos de longitud. También debían cumplir otra condición: la longitud en aminoácidos del alineamiento sólo puede diferir en 10 aminoácidos como máximo de la longitud en aminoácidos de la primera proteína de las dos comparadas. Obtuvimos estos datos consultando la tabla 'homology' de la base de datos Bionemo.

## Movilidad y organización genética

Para realizar la simulación de la transferencia de genes realizamos los siguientes pasos:

• Seleccionamos los operones que codifican complejos enzimáticos

• Obtenemos un inicio de transferencia para los genes en los operones. Para que la selección del inicio se ajustara más a la realidad, en la medida de lo posible, decidimos que el inicio de transferencia se podía encontrar en genes que estuvieran más allá del extremo 5' del operón. Así pues, el método que seguimos fue el siguiente:

• Obtenemos la longitud en genes del operón (por ejemplo, 6)

•Elegimos un número al azar entre el valor negativo de la longitud del operón y la longitud del operón sin incluir el 0 (por ejemplo, un número entre -6 y 6 que no sea 0). Para obtener este número al azar generamos una distribución uniforme de números enteros que van desde el valor negativo de la longitud del operón al valor positivo de la longitud del operón utilizando el método 'random_uniform' del módulo de Perl Math::Random.

•Si el número elegido para el inicio de la transcripción es negativo entendemos que es el número de genes más allá del extremo 5' del operón a partir del cual empezamos a contar. (por ejemplo, si elegimos el número -3 empezamos a contar:-3, -2, -1, 1, etc.)

• Seleccionamos una longitud en genes para el fragmento de ADN que vamos a transferir. Obtenemos la longitud del fragmento escogiendo números al azar que obtenemos de una distribución. Como desconocemos los mecanismos moléculares exactos de la transferencia de genes no podemos asegurar que la longitud de los fragmentos siga una distribución determinada en la realidad. Por ello probamos tres distribuciones distintas: uniforme, exponencial y normal. Para obtener los números al azar a partir de una distribución normal utilizamos el método 'random_normal' del módulo de Perl Math::Random. Elegimos un valor de media de 4 y un valor de varianza de 1,4 . Nos decidimos por estos valores porque son los que nos generaban una mayor variedad de longitudes con una media cercana a 6, que es la longitud media de los operones, y que producían pocos valores negativos de longitud que no nos interesan por razones obvias.

• A continuación escogemos el fragmento del operón que empieza en el valor elegido par el inicio de transcripción y que contiene el número de genes generado con la distribución elegida para obtener la longitud de genes (por ejemplo, si el inicio de transcripción es 2 y la longitud es 3 cogemos los genes del operón que están en las posiciones 2,3 y 4). En el caso de el operón tuviera una longitud menor de la seleccionada se descartan los genes sobrantes (por ejemplo, si el operón tiene una longitud de 3 genes, el inicio es el tercer gen y la longitud es 2 únicamente cogemos el gen 3). Si el inicio de transcripción es negativo y la longitud seleccionada no nos permite coger ningún gen del operón volvemos a repetir el evento sin tenerlo en cuenta para la suma total de eventos simulados (por ejemplo, si el inicio es -6 y la longitud es 2).

- Finalmente, calculamos el número de complejos completos codificados en los genes transferidos y lo dividimos por el número de posibles complejos completos transferidos por el operón completo. De este modo obtenemos un coeficiente de 'transferibilidad' que va de 0 (ningún complejo a transferido) a 1 (todos los complejos posibles transferidos)

- Para calcular las conexiones entre complejos transferidos seguimos el mismo procedimiento que para calcular los complejos transferidos excepto por dos salvedades:
  - Solo hicimos la simulación con operones que contuvieran al menos dos complejos conectados ya que si no no es posible que ese operón transfiera ninguna conexión
  - Para calcular el número de conexiones transferidas dividimos el número de conexiones transferidas entre el número total de conexiones posibles. De nuevo obtenemos un coeficiente de 'transferibilidad' que en este caso va de 0 (ninguna conexión a transferido) a 1 (todos las posibles conexiones transferidas)

Realizamos cada simulación 10000 veces y obtenemos una media de 'transferibilidad' para cada operón, sumando los coeficientes de transferibilidad obtenidos en cada evento y dividiéndolo por 10000, que tiene un valor que va de 0 a 1.

Para comprobar si el orden real es mejor que un orden al azar realizamos este mismo experimento probando 1000 ordenes distintos para cada operón, tanto en la simulación de transferencia de complejos como en la de transferencias de complejos conectados. Elegimos 1000 ordenes porque así podemos incluir todos los ordenes distintos de los operones que contienen 6 genes (6!) que es la media de longitud de los operones de biodegradación. También nos permite cubrir una proporción de todos los ordenes en el 84% del total de operones analizados. No elegimos una mayor número de ordenes por limitaciones de computación que hacían los cálculos muy lentos.
A continuación, calculamos una media de 'transferibilidad' para cada uno de los ordenes alternativos de cada operón sumando los coeficientes de 'transferibilidad' de cada evento y dividiendo por 10000 que es el número de veces que repetimos el experimento. Con las medias de los 1000 ordenes generamos una distribución de 'transferibilidad' para los ordenes al azar de ese operón.
Finalmente, para cada operón, calculamos un valor z para la media de 'transferibilidad' obtenida por medio de la simulación con el orden real del operón frente a la distribución de medias obtenida con los ordenes al azar de ese operón.

## Respuesta a nuevas señales, especificidad y coevolución entre regulación y metabolismo

En este apartado de nuevo utilizamos el API y la base de datos local de *Escherichia coli* para realizar los cálculos y el software R para el análisis estadístico (Ihaka y Gentleman, 1996).

Las pequeñas moléculas orgánicas desempeñan un papel fundamental en bioquímica, medicina y biología. Para poder realizar comparaciones entre grupos a gran escala y a nivel computacional se han desarrollado representaciones moleculares que se conocen como 'fingerprints' (Baldi *et al.*, 2007). Estas 'fingerprints' codifican la estructura de los compuestos químicos por medio de un vector que representa la presencia o ausencia, o el número de ocurrencias, de determinadas estructuras moleculares.

Para compararlas se utilizan coeficientes de similitud, de los cuales el más popular es el de Tanimoto (Holliday *et al.*, 2002) que se calcula dividiendo la intersección de los 'fingerprints' por la unión. De este modo dos compuestos idénticos tendrían un coeficiente de Tanimoto de 1 y dos compuestos totalmente diferentes tendrían un coeficiente de 0 (Baldi y Benz, 2008).

Para calcular las similitudes entre compuestos químicos empleamos el 'coeficiente de Tanimoto' aplicado a 'fingerprints' (Baldi y Benz, 2008). Las 'fingerprints' utilizadas fueron las ECFP_6 (Extended Connectivity Fingerprints, con 6 como máximo de diámetro empleado). El software utilizado para calcular las similitudes es 'SciTegic Pipeline Pilot version 7.0.1.100' de Accelrys Software INC.

Para calcular la homología entre operones seguimos los mismos métodos descritos en la sección de integración con la fisiología.

Para calcular la homología entre reguladores utilizamos la tabla 'homology' de la base de datos Bionemo consultando el campo 'percentage identities' que describe el porcentage de identidad entre los aminoácidos de la zona alineada entre dos proteínas.

# RESULTADOS

## Conceptos sobre regulación transcripcional en biodegradación

El control de la expresión de rutas catabólicas para la degradación de compuestos aromáticos puede ser ejercido por una gran variedad de proteínas reguladoras pertenecientes a distintas familias de reguladores. El estudio de la regulación transcripcional de estas rutas se ha limitado tradicionalmente a experimentos que describen los mecanismos que controlan una ruta en un microorganismo. Pero tanto la aparición de revisiones recopilando información sobre estos circuitos de regulación (Díaz y Prieto, 2000; Tropel y Van der Meer,2004) como la continua aparición de nuevos artículos que describen nuevos sistemas nos ha permitido acumular la suficiente información como para realizar un estudio sistemático de los circuitos de regulación.

Es un hecho conocido que los genes catabólicos que codifican los componentes de enzimas implicadas en procesos de biodegradación se encuentran frecuentemente embebidos en elementos móviles que facilitan su transferencia entre diferentes especies de bacterias por medio de procesos de conjugación (Springael y Top, 2004). La presencia de sistemas homólogos para la degradación de compuestos similares en diferentes especies también sugiere la existencia de procesos de transferencia horizontal de genes (Top y Springael,2003).

Por otro lado, se ha sugerido que las bacterias implicadas en biodegradación adquieren la capacidad de responder a nuevas señales partiendo de un situación en la que los reguladores y promotores son poco específicos. Esta falta de especificidad podría permitir reclutar reguladores para controlar la expresión de nuevos operones catabólicos que se pueden haber adquirido por transferencia horizontal de genes (HGT).

Utilizando los datos acumulados intentaremos responder a estas preguntas: ¿cómo funcionan los sistemas y qué los hace diferentes de otros? ¿cómo ha afectado la movilidad a los sistemas y su organización genética? y ¿qué influencia tiene la necesidad de integrar nuevas señales en las características de los sistemas?

# Definición de un modelo molecular para la regulación de la transcripción

Para explicar cómo funcionan los mecanismos de regulación debemos conocer primero las características de los elementos que intervienen en el proceso. La definición de unos componentes de los sistemas y de las relaciones entre ellos nos permitirá posteriormente analizar los mecanismos de regulación de forma sistemática y hacer comparaciones entre componentes y sistemas. En la figura 9 se pueden observar los distintos componentes que participan en el control de la expresión de un promotor y cuyas relaciones se explican a continuación.

Definimos operón como un conjunto de genes que se expresan desde un promotor. Esta definición implica que el mismo grupo de genes expresado desde un promotor distinto constituye un operón distinto y viceversa: el mismo promotor expresando un grupo distinto de genes también constituye un operón distinto.

Un complejo de unión esta formado por un complejo regulador y un sitio de unión al ADN. Un complejo regulador esta a su vez constituido por un grupo de proteínas reguladoras y, en ocasiones, una molécula efectora que interactúa con las proteínas.

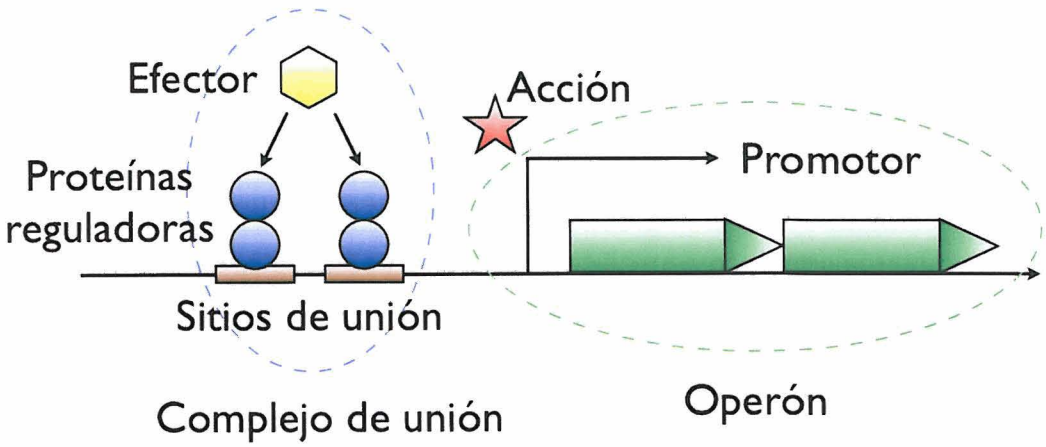



**Figura 9**. Modelo molecular de un sistema de regulación de la transcripción. Los complejos de unión ejercen una acción sobre el promotor alterando la expresión de los genes contenidos en el operón

Definimos acciones como cambios en la expresión de un promotor causados por cambios a su vez en uno o varios complejos de unión asociados al promotor. Estas acciones pueden darse tanto por la unión de un complejo regulador a un sitio de unión como por la liberación del sitio de unión por parte del complejo regulador. El resultado de las acciones puede ser un aumento de la expresión de los genes desde ese promotor, que llamaremos inducción, o un descenso de la tasa de expresión desde ese promotor que llamaremos inhibición.

# Base de datos Bionemo y sus herramientas

## Base de datos

Para poder almacenar, gestionar y relacionar la información disponible sobre la regulación transcripcional de los procesos de biodegradación construimos una base de datos relacional. Esta base de datos nos facilita el manejo de la información haciendo que sea sencillo extraer relaciones entre los distintos componentes que la forman. Todo el trabajo de construcción de la base de datos y de las herramientas para acceder a ella fue realizado junto a Almudena Trigo.

La base de datos se llama Bionemo (Biodegradation network molecular biology database) y contiene información detallada sobre la biología molecular del metabolismo y la regulación transcripcional de procesos de biodegradación. Para su implementación usamos PostgreSQL (*Berkeley Software Distribution*). La base de datos está organizada en tablas que contienen los datos y que están conectadas por relaciones entre ellas (ver detalles en Materiales y Métodos).

Las rutas metabólicas almacenadas en la base de datos son las que estaban definidas en la base de datos de biodegradación y biocatálisis de la Universidad de Minnesota en julio de 2005. Nos centraremos en la parte de regulación transcripcional que es la que trata esta tesis. También comentaremos como se combina la información metabólica con la genética y de regulación para tener una visión global de la estructura de la base de datos (figura 10).

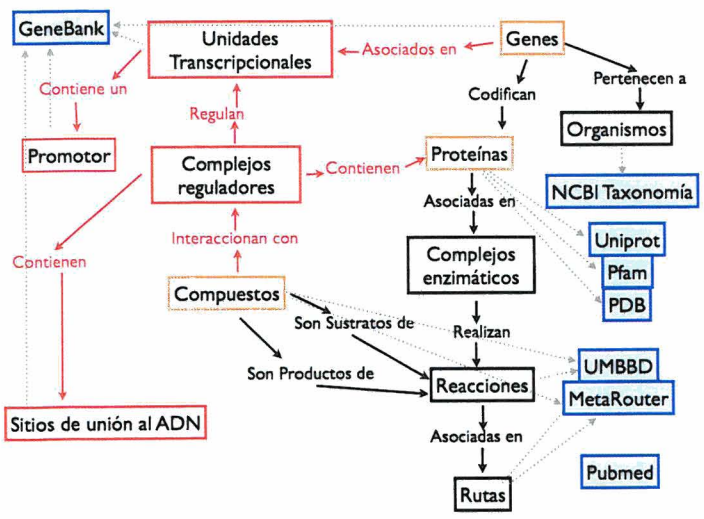



**Figura 10.** Representación esquemática de las entidades de Bionemo y sus relaciones. Las entidades biológicas contenidas en en Bionemo se muestran en las cajas de fondo blanco y las líneas sólidas indican la relación entre ellas. Las entidades implicadas en procesos de regulación transcripcional están enmarcadas en rojo y las relaciones de regulación que las conectan también están coloreadas de rojo. Las entidades que también están implicadas en procesos de regulación pero que no son estrictamente entidades de regulación están enmarcadas en naranja. Tanto los compuestos químicos como las genes codificados conectan las entidades de regulación y de metabolismo. Las cajas azules son bases de datos externas a las que diferentes entidades están conectadas por flechas grises de puntos.

La entidad central de la base de datos son los complejos enzimáticos. Estos están unidos a las reacciones bioquímicas que están definidas como transformaciones entre sustratos y productos. Las reacciones se asocian en rutas. Los complejos enzimáticos están compuestos de subunidades de proteínas codificadas por genes. Los genes están asociados en unidades transcripcionales que contienen un promotor. Estas unidades transcripcionales están reguladas por uno o más complejos de unión. Estos complejos de unión, como se ha descrito anteriormente, están formados por sitios de unión, complejos de proteínas y, generalmente, un compuesto inductor. Todas las entidades de la base de datos están conectadas con bases de datos externas que amplían la información sobre ellas.

El proceso de recopilación de la colección de datos sobre regulación y su integración con la obtenida para el metabolismo se describe en la sección de materiales y métodos. La información obtenida para las entidades de regulación se muestra en la tabla 1.

| Entidad | Cantidad |
|---:|---:|
| Organismos | 70 |
| Proteínas reguladoras | 110 |
| Complejos reguladores | 193 |
| Efectores | 88 |
| Acciones de regulación | 346 |
| Promotores | 100 |
| Unidades de transcripción | 212 |
| Genes regulados | 609 |
| Reacciones reguladas | 108 |
| Rutas reguladas | 30 |
| Compuestos degradados | 104 |

**Tabla 1.** La tabla representa la cantidad de información contenida para los distintas entidades de regulación en Bionemo. La mayoría de las categorías no requieren explicación menos la última que se refiere a que disponemos de información sobre la regulación de la degradación de 104 compuestos diferentes

## Interfaz de programación de aplicaciones (API)

Gracias a la estructura de la base de datos se pueden establecer relaciones entre entidades y al analizar los datos contenidos en ellas descubrir características de los sistemas de biodegradación difícilmente discernibles desde otra aproximación a los datos. Pero esta misma propiedad hace que el manejo de la base de datos sea complejo si no se tiene experiencia en el manejo del lenguaje SQL y si no se conoce perfectamente tanto las características de las entidades como las relaciones entre ellas. Para solucionar este problema creamos por un lado un interfaz de programación de aplicaciones (API de sus siglas en inglés) utilizando el lenguaje de programación Perl y, por otro, una interfaz Web.

Una API (del inglés *Application Program Interface*) es el conjunto de funciones y procedimientos que ofrece una cierta librería para ser utilizado por otro software como una capa de abstracción. Uno de los principales propósitos de un API consiste en proporcionar un conjunto de funciones de uso general de forma que los usuarios (programadores) puedan hacer uso de su funcionalidad evitándose el tener que programar todo desde el principio.

En nuestro caso, la API que hemos desarrollado comprende un conjunto de paquetes que se corresponden con entidades de la base de datos: organismos, ADN , genes, proteínas, complejos, unidades de transcripción, promotores, etc. Cada uno de los paquetes contiene métodos que permiten acceder a propiedades de la entidad o acceder a otras entidades relacionadas. Enlazando estos métodos es posible realizar consultas complejas utilizando un código de programación lógico y estructurado.

Como ejemplo para ilustrar la utilidad de disponer de una API mostramos como buscar las reacciones inducidas por benzoato usando SQL y usando la API. Usando SQL la consulta sería la que se muestra en la siguiente página.

```
SELECT DISTINCT r.* from reaction r LEFT JOIN reaction_complex rc ON
        r.id_reaction=rc.id_reaction LEFT JOIN complex c ON
     rc.id_complex=c.id_complex LEFT JOIN complex_protein cp ON
         cp.id_complex=c.id_complex LEFT JOIN protein p ON
p.id_protein=cp.id_protein LEFT JOIN gene g ON g.id_gene=p.id_gene LEFT
   JOIN gene_operon go ON go.id_gene=g.id_gene LEFT JOIN operon o ON
         o.id_operon=go.id_operon LEFT JOIN promoter pr ON
         pr.id_promoter=o.id_promoter LEFT JOIN action a ON
     a.id_promoter=pr.id_promoter LEFT JOIN action_b_complex ab ON
       a.id_action=ab.id_action LEFT JOIN binding_complex bc ON
      bc.id_bind_complex=ab.id_bind_complex LEFT JOIN complex ON
  complex.id_complex=bc.id_complex LEFT JOIN complex_substrate cs ON
      cs.id_complex=complex.id_complex LEFT JOIN substrate s ON
      s.id_substrate=cs.id_substrate WnERE s.name='Benzoate' and
         Sa.action='activation' or a.action='derepression');
```

Usando la API la consulta sería así:

```
get_compoundS=description'e'Benzoate')F>its_reactions_asInducer
```

El acceso a la base de datos resulta mucho más intuitivo de esta manera, no precisa de un conocimiento exhaustivo de las tablas implicadas en la consulta ni de sus campos. Simplemente consultando la documentación del API para ver las propiedades de la entidad y sus conexiones con otras se puede acceder fácilmente a la información.

## Servidor Web

Para hacer accesible toda la información a la comunidad científica, sin que sea necesario saber programar, desarrollamos un servidor Web. Este servidor permite navegar a través de Internet por las diferentes entidades contenidas en Bionemo consultando los datos

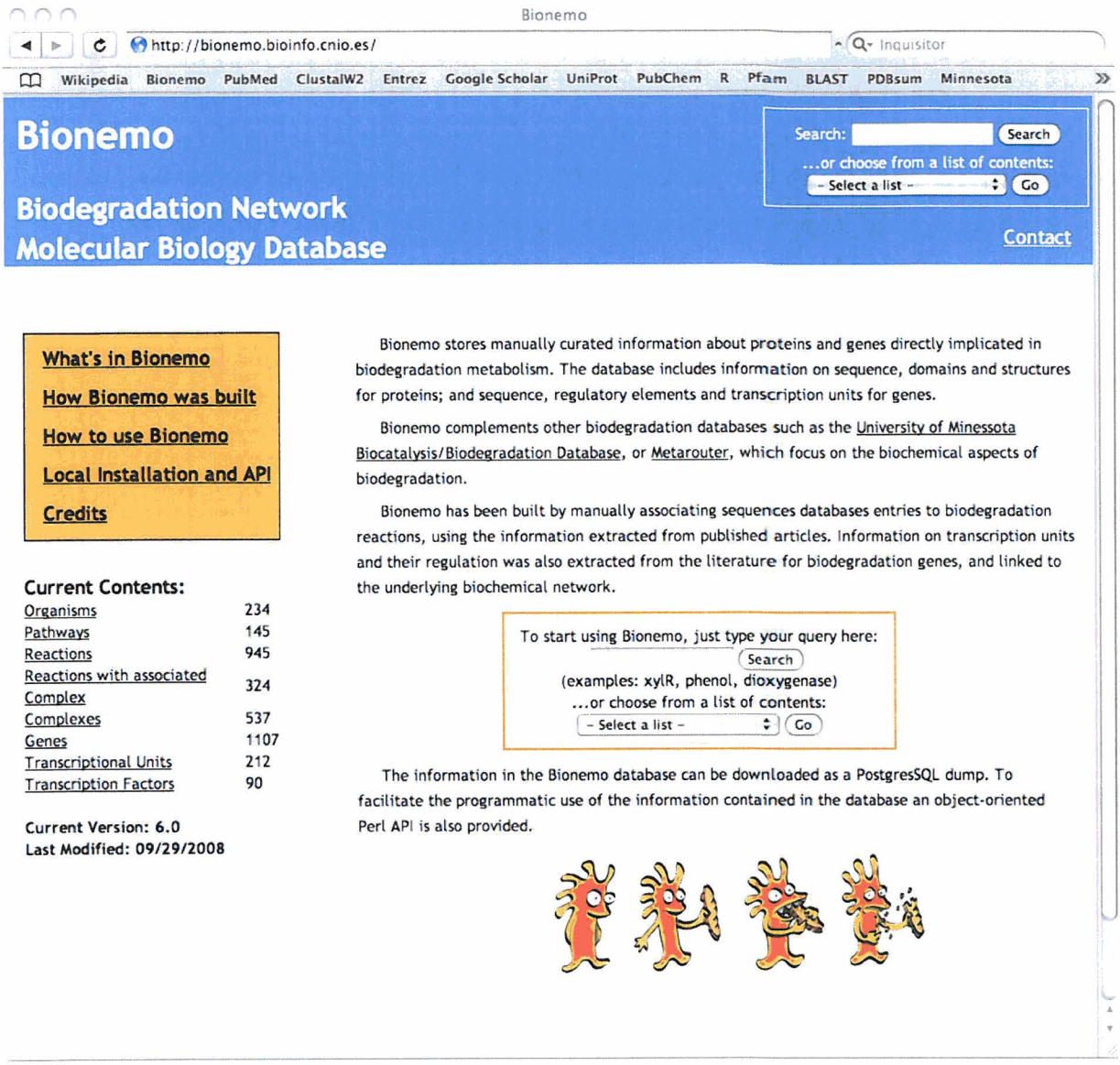



**Figura 11.** Página de inicio del servidor Web de Bionemo

contenidos sobre cada una de ellas y sus conexiones con otras entidades. Se accede a él a través de su sitio web (http://bionemo.bioinfo.cnio.es). La página de inicio (figura 11) muestra a la izquierda una lista con las diferentes entidades y la cantidad de información disponible sobre cada una de ellas.

Se puede hacer click en cada una de ellas para navegar por las listas de cada entidad pero también existe un interfaz de búsqueda que permite preguntar por contenidos disponibles en las distintas entidades. Por ejemplo, si introducimos 'benzoate' y hacemos click en 'Search ' llegaremos a una página donde se nos muestran los resultados en diferentes pestañas representando cada pestaña una de las entidades donde se ha encontrado la búsqueda. En este caso habría encontrado reacciones, complejos, rutas y compuestos (figura 12).

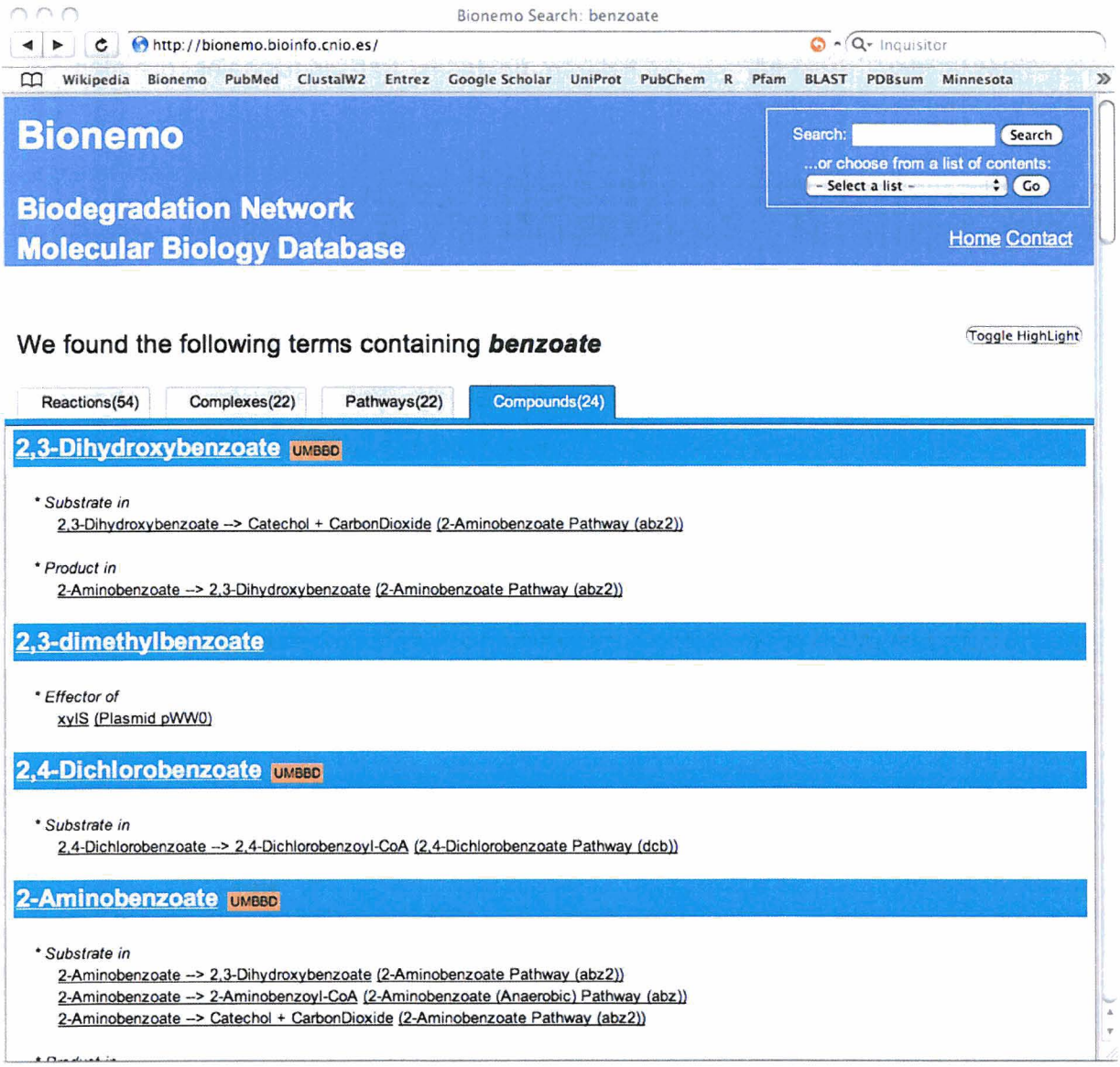



**Figura 12.** Resultado de la búsqueda 'benzoate' en el servidor web

Si miramos en la pestaña 'Compounds ' vemos que la búsqueda a encontrado todos los compuestos que contienen benzoate como parte de su nombre, como '2-Aminobenzoate' o '2,3-dimethylbenzoate'. Se puede ver que, tanto en esta pestaña como en la de rutas y

en la de reacciones existen enlaces a la página que amplía la información disponible en Bionemo con la contenida en la base de datos de la Universidad de Minnesota. Si nos fijamos en la entrada para el '2,3-dimethylbenzoate ' podemos leer ' Effector of XylS (Plasmid pWW0)'. Podemos hacer click sobre XylS e iremos a la página que describe la información del regulador (figura 13), uno de cuyos inductores es '2,3-dimethylbenzoate' como ya sabemos. En esta página encontramos en el primer bloque de información datos sobre la unidad, en este caso unidades, transcripcional que expresa el gen *xylS* y las proteínas reguladoras que controlan su expresión, XylR, IHF y HU. Conviene recordar que, según nuestra definición, dos unidades transcripcionales que contengan los mismos genes pero distinto promotor se consideran unidades transcripcionales distintas. En

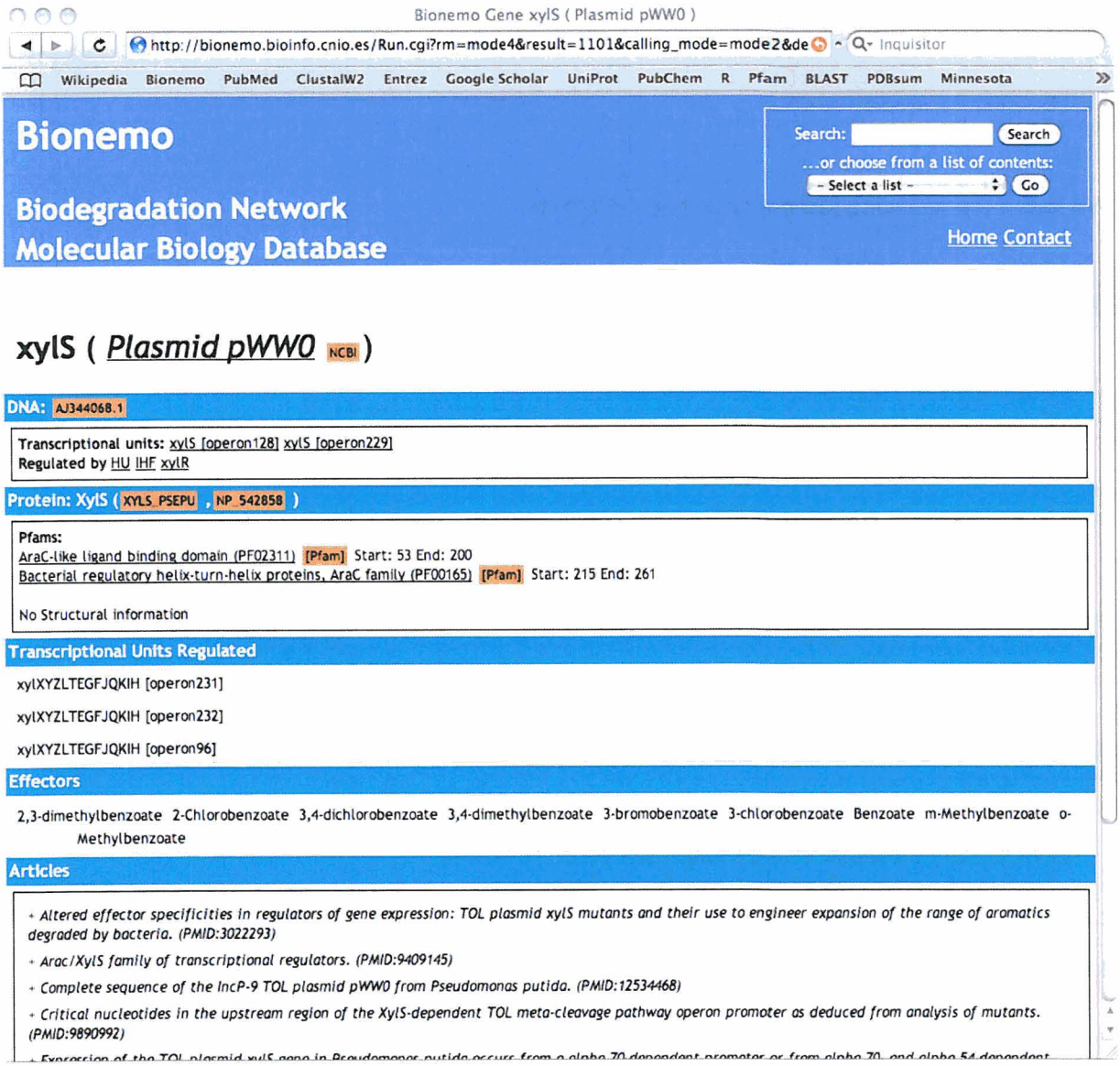



**Figura 13.** Página del regulador XylS en Bionemo

nuestro ejemplo tenemos dos unidades transcripcionales que expresan el mismo gen, *xylS*, pero desde dos promotores distintos. Continuamos con la descripción del primer bloque. Este bloque contiene información relacionada con el ADN y la expresión del gen, por ello dispone también de un enlace a GenBank donde podemos acceder a la secuencia de ADN del gen y a su contexto genómico. En el siguiente bloque se expone la información disponible asociada a la proteína (enlaces a Uniprot y Genpept, en todos los casos, y dominios de Pfam y PDBs, en el caso de que existiera esta información disponible). El tercer bloque nos muestra las unidades transcripcionales reguladas por este regulador. El cuarto muestra una lista de compuestos que actúan como efectores de XylS. Si hacemos click sobre alguno de ellos iremos a una página similar a la de la figura 12 con los resultados de la búsqueda en Bionemo de la información disponible para ese compuesto. De la misma manera, si hacemos click sobre el nombre del organismo al lado del nombre del gen en el encabezamiento de la página haremos una búsqueda en Bionemo de toda la información disponible sobre ese organismo. Finalmente, el último bloque muestra una lista de artículos que son la fuente original de la que se extrajeron los datos que pueblan la página. Se puede acceder a la página de Pubmed donde se muestra el sumario del artículo haciendo click sobre el título.

Pero, para continuar descubriendo la información sobre regulación contenida en Bionemo es mejor que hagamos click sobre una de las unidades de transcripción reguladas por XylS para acceder a la página que describe una unidad de transcripción (figura 14, siguiente página). Esta página también está divida en bloques. En el primero se muestra la información relacionada con el promotor: enlace a la entrada de GenBank donde se contiene la secuencia de ADN, posición del inicio de transcripción en esa entrada, orientación de la transcripción y factor sigma asociado a ese promotor. En el siguiente bloque se muestra la lista de genes contenidos en el operón, pudiéndose hacer click sobre ellos para ir a la página del gen. El tercer bloque muestra los sitios de unión al ADN, con su secuencia y coordenadas en un archivo GenBank y los reguladores que se unen a ese sitio de unión. El cuarto bloque muestra los detalles de la regulación con un sub-bloque para cada regulador que controle la unidad transcripcional. En esta caso es sólo XylS. Se muestra su modo de acción, activador en este caso, de nuevo los sitios de unión al ADN en coordenadas y la lista de compuestos que pueden actuar como inductores de ese regulador. También se muestra un dibujo del operón en el que se muestra la escala en pares de bases y en el que se puede hacer click sobre los reguladores y genes para ir a sus páginas. Si, por ejemplo vamos a la página de XylX el resultado es parecido al de la figura 13, ya que también es una página de un gen como XylS, pero con diferencias ya

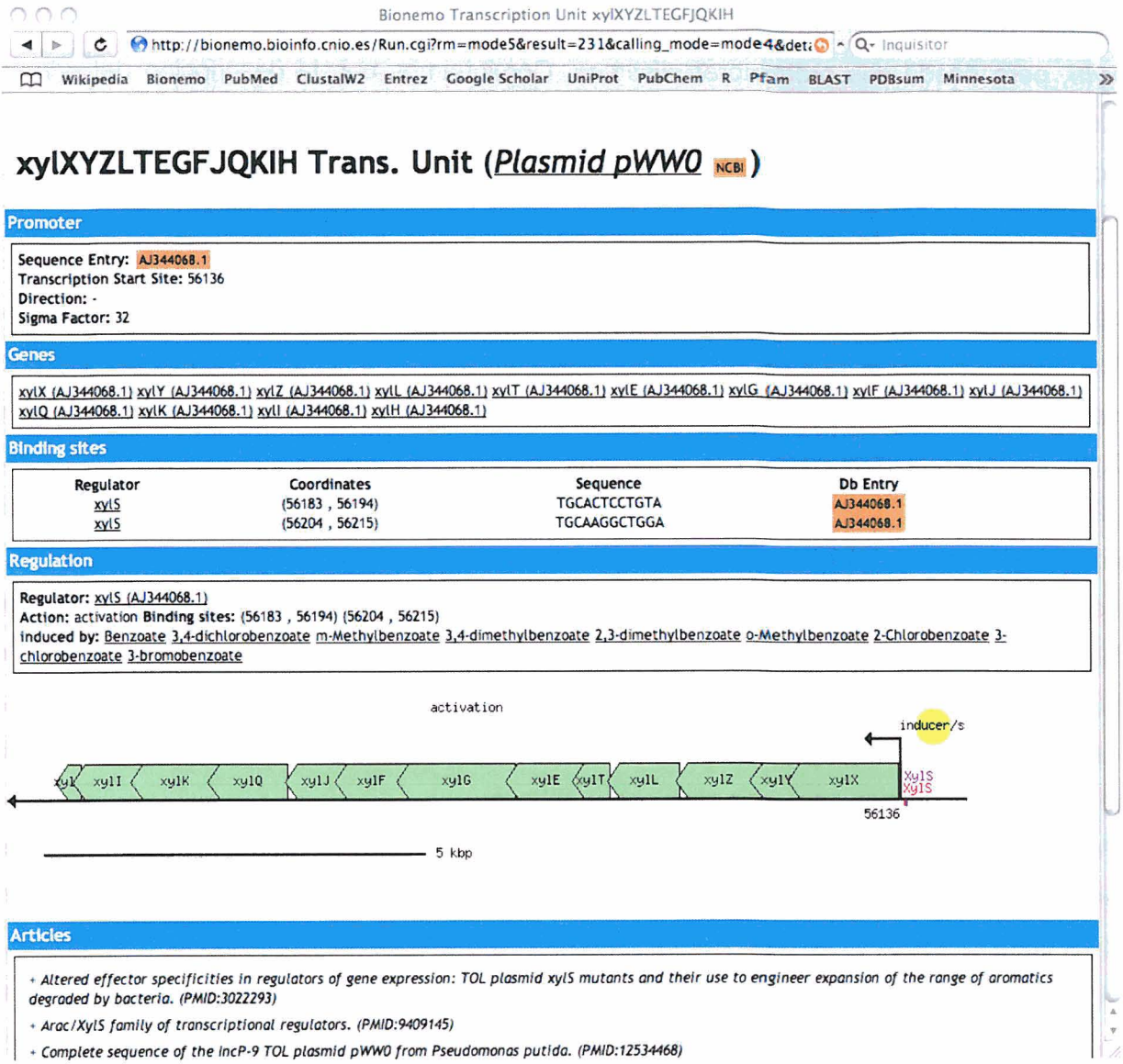



**Figura 14.** Página de la unidad transcripcional *xylXYZTEGFJQKIH* expresada desde su promotor sigma-32.

que este es un gen que forma parte de un complejo enzimático. Los dos primeros bloques se mantienen igual pero el tercero contiene información sobre los complejos enzimáticos que forma la proteína XylX y las reacciones realizadas por los complejos. No vamos a explicar en detalle la información sobre el metabolismo, solamente mencionar que si hacemos click sobre el nombre de una ruta, por ejemplo la de 'p-xylene' iremos a la página que nos muestra la ruta (figura 15, siguiente página). En el caso de que dispongamos de información sobre la regulación transcripcional de la ruta existe un botón en la esquina superior izquierda de la página. Si hacemos click sobre el se nos muestra el regulador que controla la expresión de las enzimas que realizan las reacciones de la ruta unido por lineas de color a las reacciones controladas por él. Si el regulador es un

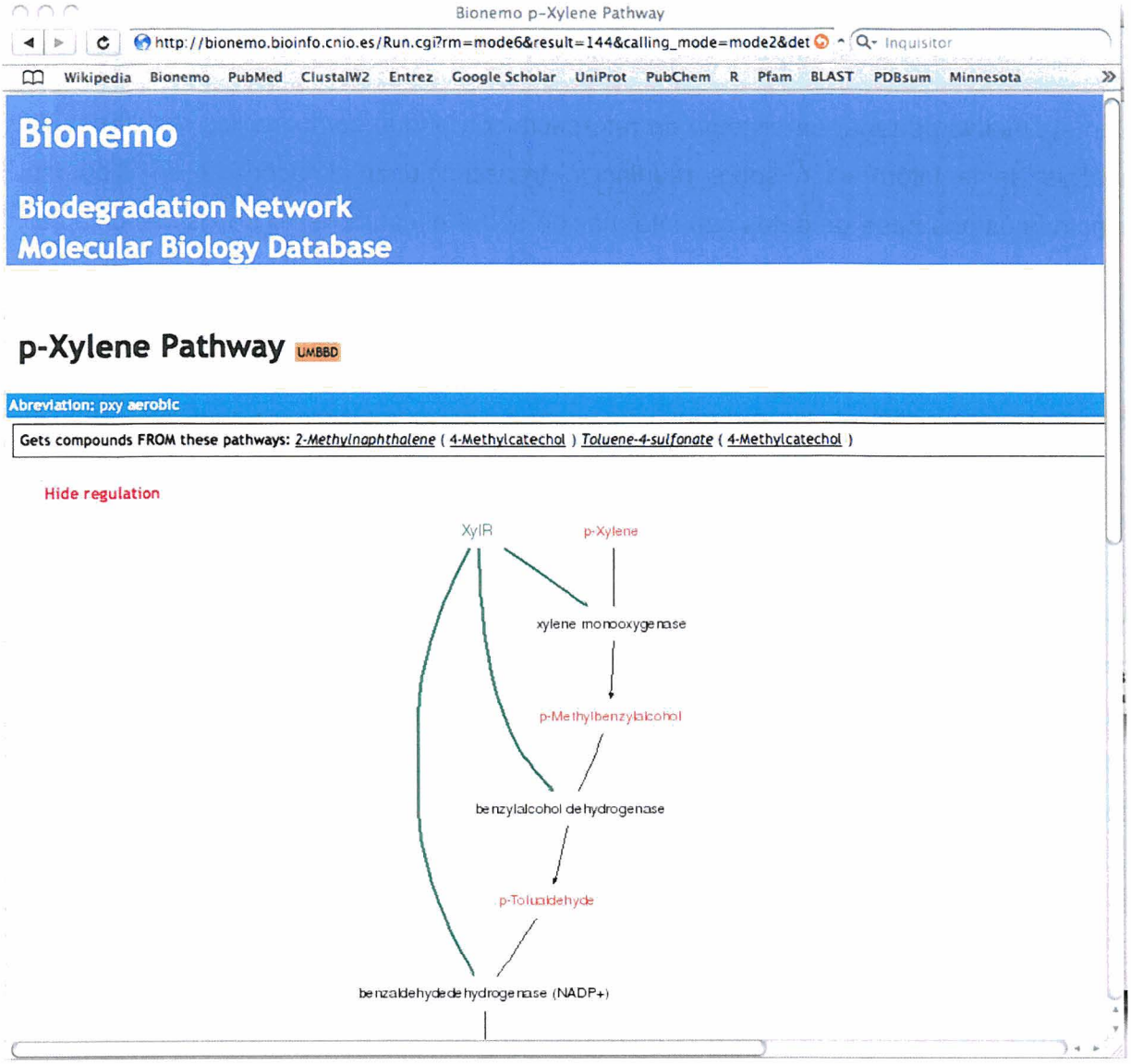



**Figura 15.** Página de la ruta del p-xileno con la información sobre los reguladores que controlan la expresión de las enzimas que realizan ciertas reacciones de la ruta.

activador, como el caso de XylR, tanto su nombre como las lineas que lo conectan con las reacciones serán de color verde. Si es un represor el color será rojo.

En resumen, tanto el servidor Web como la API hacen que la información que hemos recopilado sobre biodegradación sea fácilmente analizable y permite descubrir o comprobar relaciones entre entidades con gran facilidad.

# Base de datos del catabolismo en *Escherichia coli*

Con el objetivo de tener un sistema de referencia con el que comparar los resultados del análisis de la información sobre regulación transcripcional almacenada en Bionemo, construimos una base de datos con información sobre el catabolismo de *Escherichia coli*. La información fue extraída de Ecocyc (ver detalles en Materiales y métodos). Esta base de datos contiene toda los datos disponibles sobre las reacciones implicadas en procesos de catabolismo, los complejos enzimáticos que las realizan y los operones que contienen los genes que codifican las proteínas que forman los complejos enzimáticos. También contiene información sobre los reguladores que controlan la expresión de esos operones y todo ello almacenado siguiendo el mismo modelo molecular que utilizamos para Bionemo siempre que fue posible y en función de nuestras necesidades (ver detalles en Materiales y métodos). De esta manera se facilita la comparación entre las propiedades de los sistemas y los componentes de regulación de *Escherichia coli* y los de biodegradación. En todos los casos en que se realizan comparaciones con *Escherichia coli*, a menos que se mencione lo contrario, estas están calculadas en base a los datos contenidos en esta base de datos.

# Caracterización de los circuitos de regulación

Una vez que tenemos toda la información sobre regulación recopilada y estructurada podemos proceder a su análisis para determinar las propiedades tanto de los sistemas reguladores como de sus componentes.

**Los sitios de unión al ADN tienen una longitud en pares de bases similar en biodegradación y en *E. coli***

Empezamos a caracterizar lo que hemos llamado 'complejos de unión' analizando las propiedades de los sitios de unión al ADN. Queremos saber si los sitios de unión tienen alguna propiedad intrínseca que los haga diferentes. Para ello empezamos calculando su longitud en pares de bases y la comparamos con la calculada para los sitios de unión de *Escherichia coli* (figura 16).

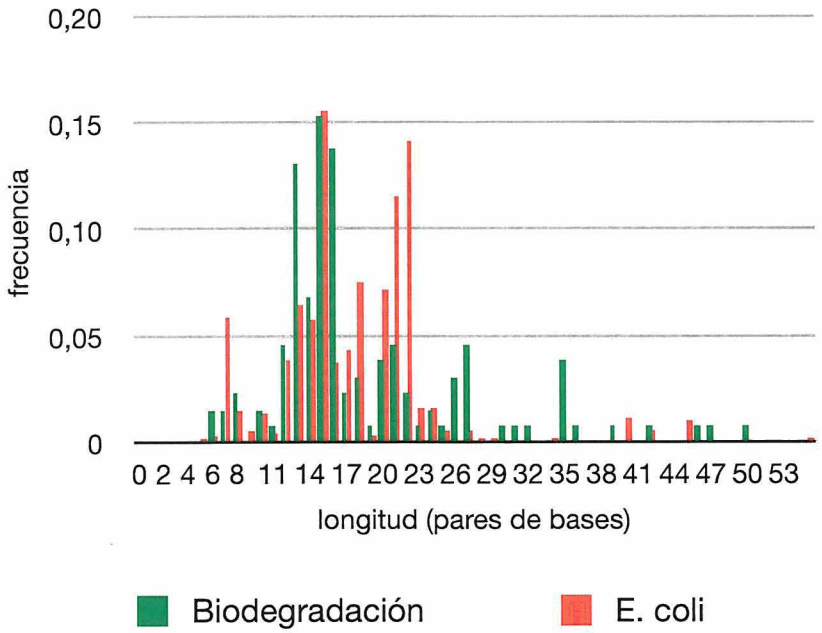



**Figura 16.** Longitud en pares de bases de los sitios de unión al ADN de los reguladores de biodegradación frente a los sitios de unión al ADN de los reguladores de las rutas catabólicas de *Escherichia coli*.

La diferencia entre las longitudes de los sitios de unión al ADN no es significativa (t-student p = 0,03). En conclusión, no parece que sea necesario que la longitud de los sitios de unión en sistemas de biodegradación sea diferente de la de los de *Escherichia coli* para que puedan realizar su función.

**Los reguladores son más frecuentemente activadores en biodegradación que en *E. coli***

Continuamos la caracterización de los 'complejos de unión' analizando las proteínas reguladoras. Se ha sugerido que la regulación de la degradación de los compuestos aromáticos está generalmente bajo el control de activadores (Díaz y Prieto, 2000). Para contrastar esta afirmación, calculamos la proporción de los diferentes tipos de reguladores en nuestra muestra y la comparamos con la proporción encontrada en *Escherichia coli* para comprobar si existen diferencias significativas entre las proporciones de los tipos de reguladores entre las dos muestras (figura 17). Hay que tener en cuenta para interpretar estos datos que clasificamos los reguladores en base a sus acciones sobre sus promotores regulados que están expresando genes catabólico, no los de genes de reguladores. De este manera, por ejemplo, si un regulador actúa como activador sobre un operón catabólico y como represor ante un gen regulador será clasificado como activador.

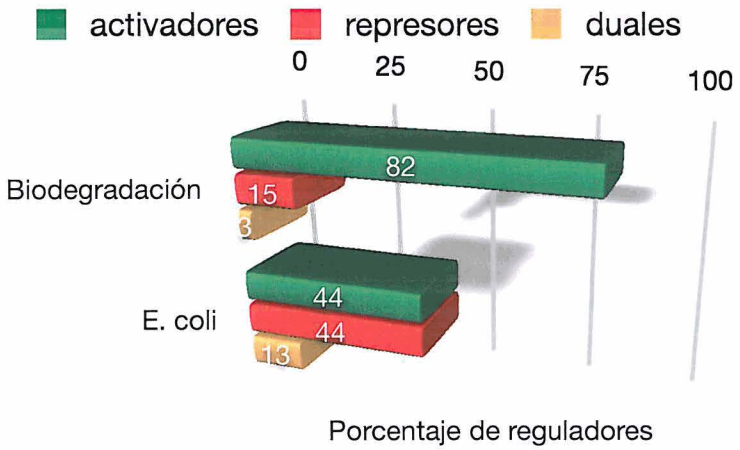



**Figura 17.** Porcentaje de tipos de reguladores según su modo de acción: activadores, represores o duales. Comparación entre los reguladores de los sistemas de biodegradación frente a los reguladores que controlan genes catabólicos en *Escherichia coli*

Efectivamente, la proporción de activadores en los sistemas de biodegradación es mucho mayor que la de cualquier otro tipo de regulador y representa un 82% del total de los reguladores frente al 44% que representa el número de activadores frente al total de reguladores en *Escherichia coli*. Existen diferencias significativas entre las proporciones de activadores de los sistemas de biodegradación y los sistemas catabólicos de *Escherichia coli* (ji-cuadrado < 0,001). Pero, ¿de donde viene está mayor proporción de activadores de los circuitos de biodegradación? Una posibilidad sería que el origen fuera un efecto fundador: cuando una población queda mermada los genes de la muestra superviviente pueden no ser representativos del acervo anterior a la merma. Este cambio en el acervo genético se llama efecto fundador, porque las poblaciones pequeñas de

organismos que invaden un territorio nuevo (fundadores) están sujetas a él. En nuestro caso de estudio esto supondría que un activador hubiera sido el primer mecanismo de regulación utilizado para el control de la expresión de genes de biodegradación y que, independientemente de que confiera una ventaja adaptativa a los microorganismos que lo tuvieran frente a los regulados por represores, su presencia se hubiera extendido. Esta hipótesis implicaría que todos los activadores tuvieran un origen común. Para contrastarla, comprobamos los orígenes de los diferentes reguladores por medio de su clasificación por familias (figura 18).

En la clasificación por familias, de nuevo, sólo hemos considerado la acción de los reguladores sobre los promotores de los genes catabólicos. A pesar de que la familia más representada es de activadores, LysR con 29 miembros, la mayor proporción de activadores no se puede atribuir solamente a la presencia de esta familia. Conviene señalar que los dos casos de reguladores duales que encontramos anteriormente se corresponden con dos reguladores de la familia LysR: BenM y CatM. Ambos reguladores son capaces de inhibir la expresión de sus operones catabólicos regulados en presencia de una gran cantidad de inductor en el medio (Chugani *et al*, 1998). La segunda familia más representada, XylR/DmpR con 17 miembros, también está formada por activadores.

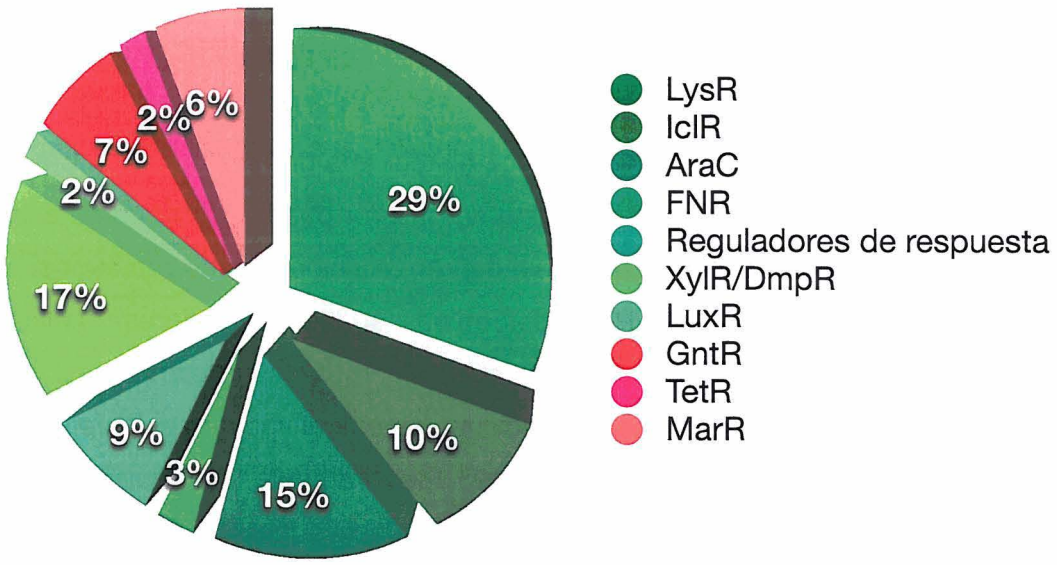



**Figura 18.** Clasificación de los reguladores de biodegradación por familias. Las diferentes tonos de azul representan diferentes familias de activadores y los diferentes tonos de rojo a diferentes familias de represores.

También la tercera familia más representada, AraC con 15 miembros, la cuarta, IclR con 10 miembros, e incluso la quinta, reguladores de respuesta con 9 miembros, son de activadores. Por otra parte, incluso en la familias de represores existen excepciones: BadR de *Rhodopseudomonas palustris* es un regulador de la familia MarR que, al

contrario que el resto de su familia, actúa como activador (Egland y Harwood, 1999) y BphR1 (también llamado Orf0) de *Pseudomonas pseudoalcaligenes* KF707 es un regulador de la familia GntR que actúa como activador (Watanabe *et al* 2003). Esto parece apoyar claramente que es un caso de convergencia evolutiva donde la regulación por activadores confiere alguna ventaja adaptativa que es seleccionada y que no nos encontramos ante un efecto fundador.

**Los reguladores pueden activar o reprimir sus promotores en función de la distancia a la que se unen al inicio de transcripción**

Para cerrar el análisis de los 'complejos de unión' estudiamos la interacción entre los reguladores y los sitios de unión al ADN. A principio de la década de los noventa se realizó un estudio por el que se relacionaba la distancia de los sitios de unión del regulador al ADN con el modo de acción de los reguladores. Se analizaron, por un lado, sitios de unión asociados a promotores sigma-70 de *Escherichia coli* y *Salmonella typhimurium,* y por otro sitios de unión asociados a promotores sigma-54 de *Escherichia coli* y *Klebsellia pneumoniae* . La posición de los sitios de unión asociados a promotores sigma-54 es en general más flexible pero en los asociados a promotores sigma-70 se encontró un patrón claro que señalaba que los represores tienden a estar más cerca del inicio de transcripción, incluso solapándolo, mientras que los activadores suelen estar más lejos. La distancia de los sitios de unión parece ejercer una gran influencia sobre la acción reguladora, hasta el extremo de que un mismo regulador puede actuar como activador si se une a una determinada distancia del inicio de transcripción y como represor si se une a otra (Collado-Vides *et al.*,1991). Para determinar si los reguladores implicados en biodegradación siguen también este patrón calculamos las distancias al inicio de transcripción de los sitios de unión y las relacionamos con el tipo de acción ejercida sobre el promotor por el complejo regulador (figura 19). Efectivamente los reguladores siguen el mismo patrón y encontramos una mayor densidad de activadores en la zona de -70 a -60 pares de base de distancia al inicio de transcripción. La mayor densidad de represores está localizada flanqueando el inicio de transcripción, especialmente en las zonas comprendidas entre -20 y -10 y 10 y 20 pares de bases de distancia del inicio de transcripción. También aquí un mismo regulador puede actuar como activador o represor en función de la distancia que lo separe del inicio de transcripción. Como ejemplo podría servir NahR, un regulador de la familia LysR que controla la expresión de los genes que degradan naftaleno del plásmido NAH7 contenido en *Pseudomonas putida* (Park *et al.*,

2005). Este regulador, por un lado, activa la expresión de los genes catabólicos uniéndose a un sitio de unión cuyo punto medio está a -64 pares de bases del inicio de transcripción. Por otro lado, el regulado reprime su propia expresión uniéndose a un sitio de unión cuyo

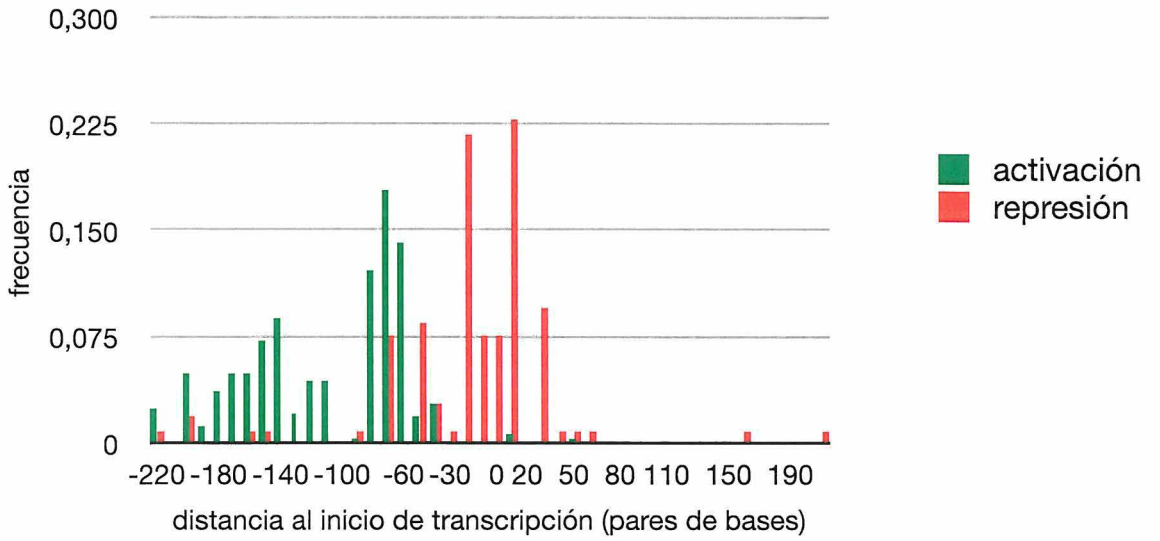



**Figura 19.** Distancia de los sitios de unión al inicio de transcripción en función del tipo de acción que realiza el regulador. La serie roja representa los sitios de unión donde el regulador actúa como un represor y los serie azul representa los sitios de unión en los que el regulador actúa como un activador. El eje de ordenadas representa la frecuencia con que un sitio de unión a la distancia representada en las abscisas pertenece a la categoría representada por la serie. Por ejemplo: entre -60 y -50 pares de bases hay 45 sitios de unión donde el regulador actúa como activador.

punto medio está 2 pares de bases por delante del inicio de transcripción. En este caso no consideramos necesario separar promotores de genes catabólicos de promotores de genes reguladores ya que el patrón de la acción de los reguladores con respecto de la distancia al inicio de transcripción es independiente del tipo de genes expresados (Collado-Vides *et al.*,1991).

## Los promotores catabólicos en biodegradación son más frecuentemente activables mientras que los de los reguladores son reprimibles

Continuamos la caracterización de los componentes implicados en la regulación transcripcional analizando los promotores desde los que se expresan los genes. Los promotores de los genes catabólicos de nuestra muestra de estudio están bajo el control de al menos un regulador, ya que ese fue nuestro criterio inicial para seleccionarlos, ya que queremos estudiar la regulación transcripcional en las bacterias que realizan procesos de biodegradación. Sin embargo, los promotores que expresan los genes de los

reguladores que controlan los promotores catabólicos no tienen porqué estar controlados por algún regulador necesariamente, de ahí que encontremos algunos constitutivos. Separamos los promotores que vamos a analizar en promotores catabólicos y promotores reguladores, según expresen enzimas o reguladores respectivamente. En el caso de que expresen tanto reguladores como enzimas los hemos incluido una vez en cada categoría analizándolos como parte de cada grupo. Por un lado, nos interesa saber cual es la diferencia que existe entre los tipos de promotores de los genes catabólicos y los de regulación. Por otro, queremos comprobar si existen diferencias con lo que encontramos en *Escherichia coli.* Primero clasificamos los promotores catabólicos según su modo de regulación en activables, reprimibles y duales (figura 20).

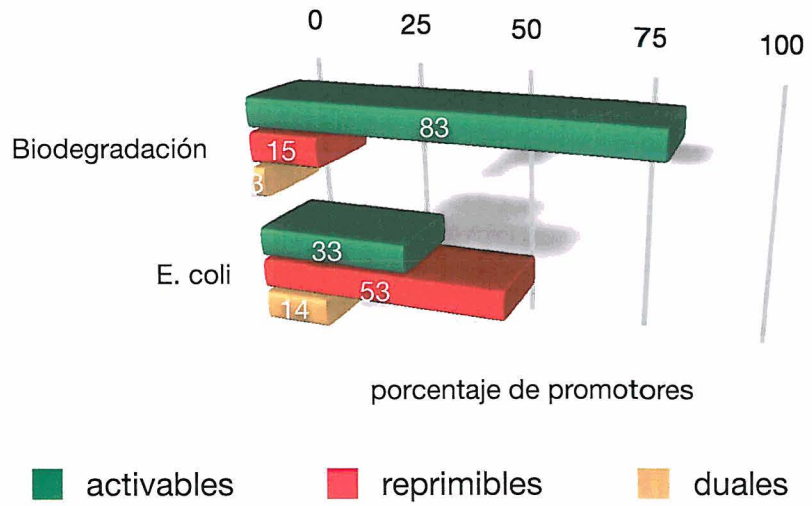



**Figura 20.** Comparación de los porcentajes de los tipos promotores que expresan enzimas en los sistemas de biodegradación y *Escherichia coli* en función de su tipo de regulación. Si están regulados por activadores son activables, si están regulados por represores son reprimibles y si están regulados por los dos tipos son duales

En el caso de los promotores de biodegradación la proporción de promotores controlados por activadores es mucho mayor que la de los promotores controlados por represores. Los promotores activables representan un 82,5% del total de los promotores mientras que los reprimibles tan sólo representan un 15%. En *Escherichia coli* la situación es la contraria ya que existe una mayor proporción de represores que de activadores. Aún así, la situación en *Escherichia coli* está mucho más equilibrada en cuanto a las proporciones de represores, que representan un 52,7% del total frente al 32,3% que representan los activadores. La proporción de promotores controlados por activadores es mayor en los sistemas de biodegradación que en *Escherichia coli* de forma significativa (ji-cuadrado, $p < 0,001$). Habría que destacar también una mayor presencia de promotores duales en

*Escherichia coli,* donde también encontramos un caso de un promotor constitutivo (no incluido en la gráfica).

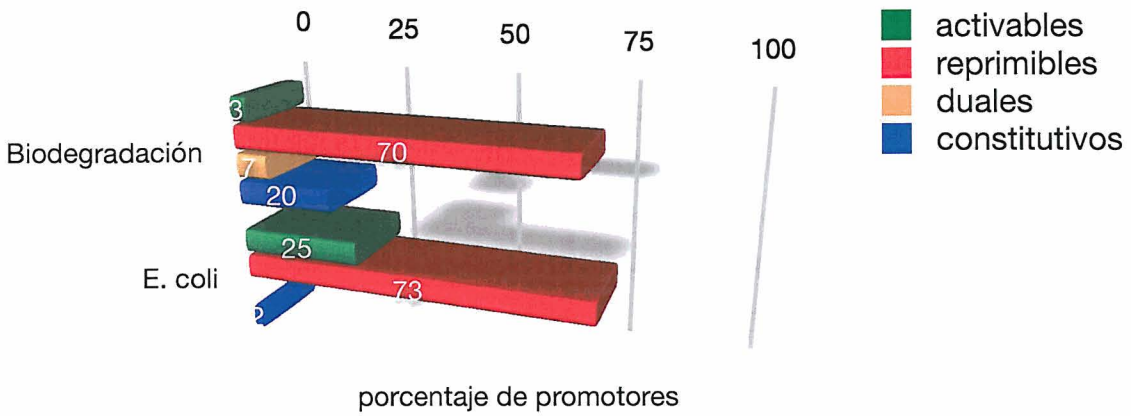



**Figura 21.** Comparación de los porcentajes de promotores que expresan operones que contienen reguladores en Bionemo y *Escherichia coli* en función de su tipo de regulación. Si están regulados por activadores son activables, si están regulados por represores son reprimibles y si están regulados por los dos tipos son duales

Pasamos a analizar lo que ocurre en los promotores de expresan genes de reguladores (figura 21). La mayoría de los promotores son reprimibles, esta vez tanto en el caso de los promotores de reguladores de biodegradación como en los promotores de reguladores del catabolismo de *Escherichia coli*. Hacemos el test de la ji-cuadrado para comparar las proporciones y obtenemos un valor p = 0,74. Esto nos indica que no se aprecian diferencias significativas entre la proporción de promotores reprimibles entre los dos grupos. De nuevo, las diferencias dentro de los grupos en las proporciones de promotores reprimibles frente a las de promotores activables están más acentuadas en el caso de los promotores de biodegradación. En este tipo de promotores el 70% son reprimibles frente al 3% de promotores activables, tan sólo un caso. En *Escherichia coli* el 73% son reprimibles y el 25% son activables. Habría que destacar el hecho de que la proporción de promotores constitutivos es significativamente mayor en biodegradación, con un valor p para el test de la ji-cuadrado menor de 0,001 (aunque esto podría deberse a falta de información).

**Los promotores catabólicos en biodegradación son frecuentemente inducibles y están controlados por activadores**

A continuación, para obtener una perspectiva global de la regulación estudiamos las acciones sobre los promotores que pueden suponer una inducción o una inhibición del promotor. Así pues, definimos circuitos de regulación formados por la pareja formada por un promotor y un regulador y, en la mayoría de las ocasiones, una molécula efectora . Se ha sugerido que los circuitos de regulación evolucionan de forma independiente a los genes que regulan (Cases y de Lorenzo, 2001). También se ha propuesto que la unidad evolutiva de control transcripcional no sea el regulador o el promotor si no el circuito que forman (Cases y de Lorenzo, 2005). Por eso nos interesa analizarlos para comprobar si hay un patrón específico de los circuitos de biodegradación que nos pueda dar pistas sobre su evolución. En el caso de la inducción tenemos dos posibilidades. Por un lado podemos tener un activador que se une al promotor en presencia de un efector induciendo su expresión. Este sería el caso, por ejemplo, de los reguladores de la familia LysR. La otra posibilidad es un represor que este bloqueando la expresión pero que libere al promotor cuando aparece una molécula efectora permitiendo su inducción. Este sería el caso de la mayoría de los genes de la familia GntR. En el primer caso tendríamos un sistema inducible por activador. En el segundo el sistema sería inducible por represor. Un promotor que sea tanto inducido como inhibido será considerado dual. Finalmente, un promotor que se expresa sin regulación es constitutivo y un promotor que inhibe su expresión por la unión de un regulador es inhibible. Comparamos las proporciones de cada sistema en biodegradación y en *Escherichia coli* (figura 22).

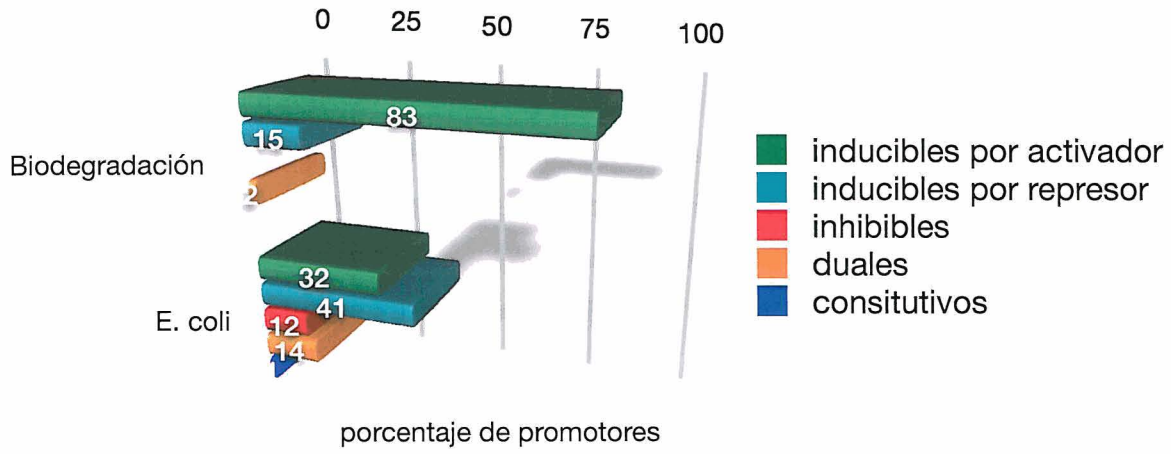



**Figura 22.** Comparación de los porcentajes de promotores según su expresión y su modo de regulación entre los operones catabólicos de biodegradación y los de *Escherichia coli*.

En biodegradación encontramos una proporción mucho mayor de promotores inducidos por activadores, que representa un 82,5% frente al escaso 15% de los inducidos por medio de represores. La situación cambia totalmente en *Escherichia coli* donde hay más inducción mediada por represor que inducción mediada por activador. También destaca el hecho de que mientras en los promotores de biodegradación hay tan sólo dos casos de promotores duales, y que por lo tanto se pueden inhibir, que representan el 2,5% del total en *Escherichia coli* hay 24 casos de promotores que se pueden inhibir y que suponen el 25,8%. Estos últimos están repartidos entre 11 puramente inhibibles, el 11,8% del total, y 13 duales, el 14%. Si miramos en detalle los casos de promotores duales de biodegradación encontramos que uno es el caso del operón *aphKLMNOPQB* de *Comamonas testosteroni* TA441 sobre el que actúan dos reguladores. Por un lado, AphR actúa como activador en respuesta a fenol. Por el otro AphS reprime la expresión del operón. El otro caso de promotor dual es totalmente distinto. En este caso el operón *pheBA* es activado por CatR que tiene dos sitios de unión para inducir la expresión del promotor y un tercer sitio de unión más allá del inicio de transcripción del promotor regulado. Se ha propuesto que CatR se une a él cuando hay gran cantidad en el medio del producto generado por las enzimas expresadas en el operón que regula y que le sirve como inductor: cis,cis-muconato. Esto permitiría inhibir la expresión de las enzimas cuando hay se han producido en cantidad suficiente lo que añadiría una capa de regulación más fina al sistema (Chugani *et al*, 1998). Si comparamos las proporciones de inducción de los genes catabólicos en biodegradación con las de *Escherichia coli* vemos que existe evidencia estadística que indica que en biodegradación la proporción de inducciones es mayor que en *Escherichia coli* (ji-cuadrado, $p < 0,001$).

**Los circuitos de regulación que controlan reguladores están generalmente formados por represores que inhiben los promotores de forma similar a *E. coli***

Pasamos ahora a analizar los sistemas de regulación de los promotores que expresan genes de reguladores (figura 23).

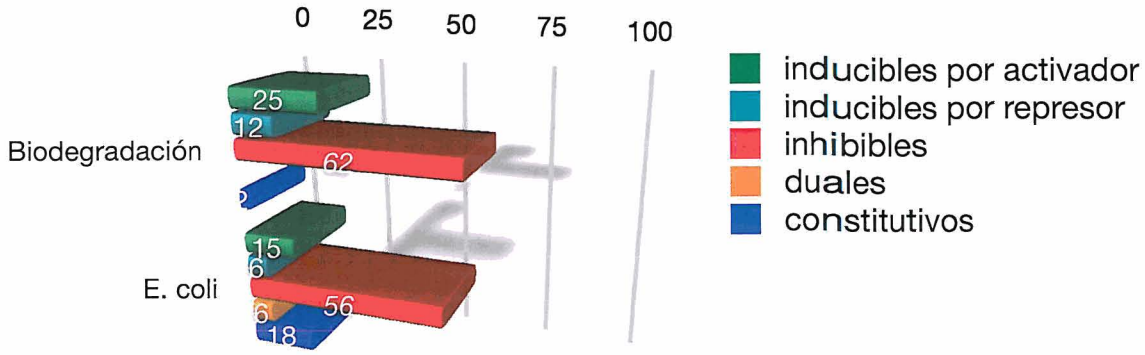



**Figura 23.** Comparación del modo de expresión de los promotores que expresan reguladores entre biodegradación y *Escherichia coli* según su expresión y su modo de regulación.

En esta ocasión no encontramos diferencias como en el caso de los promotores catabólicos ya que tanto la mayoría de los promotores de biodegradación como los de *Escherichia coli* son inhibibles. Para ver si las proporciones son similares hacemos el test de la ji-cuadrado. Obtenemos un valor p = 0,58. Esto nos indica que no existe evidencia estadística para rechazar la hipótesis de que las dos proporciones son iguales.

**Los promotores de los genes reguladores que activan los promotores catabólicos suelen reprimirse haciendo que el sistema sea más estable**

Recientemente, en un estudio sobre el diseño de los circuitos regulatorios en bacterias se propuso que la forma de tener un sistema más estable, robusto y sensible para la respuesta a un inductor depende del tipo de regulador que controla la expresión de sus genes regulados y de la relación entre el tipo de regulador y la auto-regulación de su gen. En concreto, la mejor manera de tener un circuito robusto que responda un efector es acoplar de forma inversa la expresión del promotor regulador y del promotor del regulador. Definían como robustez la capacidad de un circuito de mantenerse en equilibrio, o no cambiar de forma significativa, cuando la estructura del sistema (los valores de los parámetros) cambian de forma significativa. Por ejemplo, si la expresión del promotor regulado aumenta la del gen del regulador debería de disminuir para que el sistema se

mantenga en equilibrio. Esto, en principio se podría conseguir de dos maneras. La primera sería con un activador que en respuesta a un inductor active la expresión de un promotor e inhiba la de su propio gen que expresa la proteína reguladora. La segunda sería por medio de un represor que libere la represión de su promotor regulado y reprima la del gen que lo expresa (Wall *et al.*, 2004). Si observamos lo que ocurre en biodegradación con la regulación de los operones catabólicos y la auto-regulación del gen que expresa el regulador que los controla, vemos que en la mayoría de los casos encontramos un promotor catabólico inducido por un activador que reprime la expresión de su propio gen (figura 24 arriba). Esta situación se da en el 48,8% de los casos de

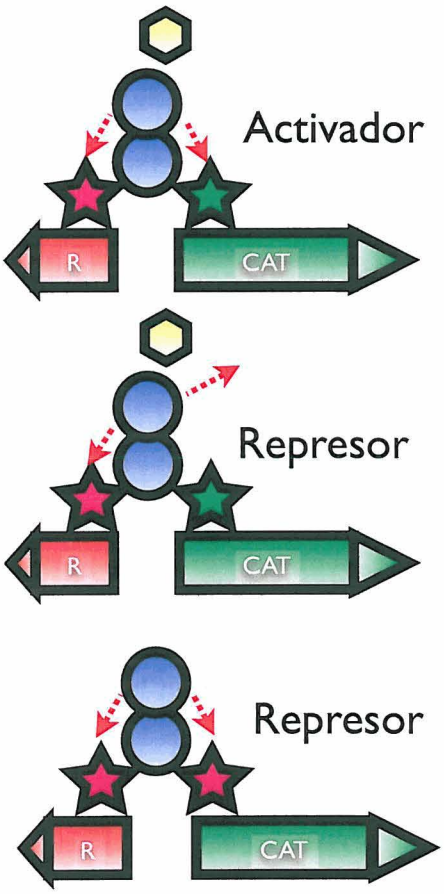



**Figura 24.** Representación de los sistemas de parejas de regulación del promotor del operón catabólico (CAT, verde) y auto-regulación del gen que expresa el regulador (R, rojo). El primer sistema (arriba) representa un activador que en respuesta a un inductor (hexágono amarillo) induce (estrella verde) el promotor catabólico y que reprime (estrella rosa) el promotor que expresa el propio gen del regulador. Las flechas rojas punteadas hacia abajo representan que el regulador se une a al ADN para realizar la acción representada por la estrella. El segundo sistema (centro) representa un sistema con el mismo resultado a nivel de expresión de los promotores pero mediado por un represor en lugar de por un activador. La única diferencia es que el represor libera el sitio de unión al ADN (representado por la flecha roja punteada hacia arriba) para permitir la inducción del promotor catabólico. El tercer sistema representa un represor que inhibe tanto el gen catabólico como el gen del regulador al unirse al ADN

biodegradación para los que tenemos información tanto de la regulación del promotor catabólico como del que expresa el regulador. En comparación en *Escherichia coli* tan sólo hemos encontrado esta situación en el 8% de los casos. La segunda posibilidad de inducir un promotor auto-reprimiendo la expresión del regulador pasa por utilizar un represor que libere su expresión. Esta situación representa el 17% de los casos de biodegradación y en *E. coli* no se da. Curiosamente, la represión de tanto el promotor catabólico como el del promotor que expresa el regulador que reprime el operón catabólico sólo se da en una ocasión en biodegradación, representando el 2,4% de los

casos, pero es la que más representada está en *E. coli,* con 13 casos que suponen un 26% del total .

Se podría argumentar que los sistemas de biodegradación son sistemas más robustos por si mismos que los de *Escherichia coli*. Puede que sea así por estar estos últimos más integrados en la célula mientras que los de biodegradación son más móviles y pasan de unos organismos a otros habiéndose seleccionado por ello los que permanecen más estables.

**El operón del regulador y su operón regulado se encuentran frecuentemente contiguos en el ADN y transcritos de forma divergente**

Una vez habíamos analizado la relación lógica entre los elementos que componen los circuitos de regulación nos preguntamos si existiría una relación física asociada. En otras palabras, ¿cómo se organizan los operones catabólicos y reguladores en el genoma?. Se ha observado en *Escherichia coli* que los operones co-regulados o que expresan genes que controlan otros operones suelen estar cerca entre sí (Warren y Wolde , 2004), así como aquellos que actúan en respuesta a estímulos externos (Janga *et al.*, 2007). En ambos estudios se proponía que esta disposición podría favorecer la co-regulación de los operones. Al pertenecer nuestros operones estudiados a estas dos categorías nos preguntábamos si también seguirían este mismo patrón. Para comprobarlo calculamos la la cantidad de genes codificados en el ADN entre el gen del regulador y el operón regulado. A continuación clasificamos como contiguos a los que no tenían ningún gen entre el gene del regulador y el operón regulado y como separados a los que sí (figura 25).

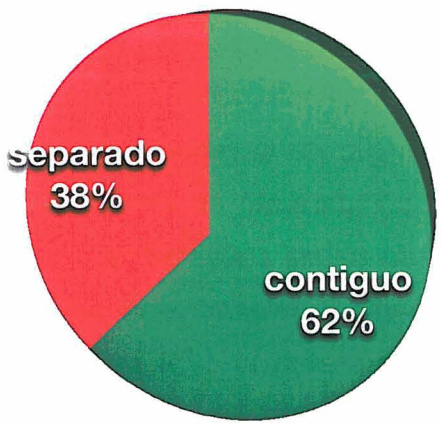



**Figura 25**. Posición del gen del regulador con respecto al operón regulado. La figura muestra la proporción de operones que tienen el gen del regulador que controla al operón contiguo (sector verde, marcado como contiguo) frente a la proporción de operones que tienen algún gen entre el operón y el gen del regulador que los controla (sector rojo, marcado como separado)

En la figura 25 se observa que la mayoría de los operones tienen el gen del regulador que los controla contiguo. La proporción de operones catabólicos contiguos al gen del regulador representa el 62% de los casos, 36 de 58 analizados, mientras que los que no tienen el gen del regulador contiguo representan el 38% de los casos, 22 de 58. Incluso entre los que no están contiguos las distancias son pequeñas ya que 8 de los 22 están a menos de 5 genes de distancia y tan sólo 2 están a más de 12 genes de distancia. Una razón que podría explicar está organización genética sería que se haya seleccionado una estructura compacta que permite codificar un circuito regulatorio con su operón regulado en un fragmento de ADN más pequeño.

Como ya hemos comentado, en los estudios anteriormente citados se menciona que los operones regulados y el gen del regulador cuando están contiguos suelen transcribirse de forma divergente lo que facilita la co-regulación al compartir la zona de unión de los reguladores al ADN (Warren y Wolde , 2004; Janga *et al.*, 2007). Comprobamos esto en nuestros sistemas comparando lo que ocurre en los operones con el gen del regulador contiguo frente a los que tienen algún gen entre los dos operones. El resultado se muestra en la figura 26.

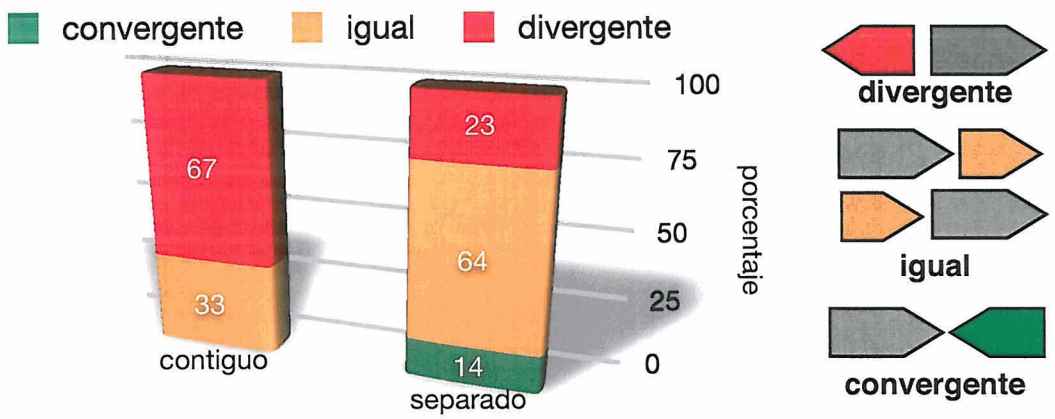



**Figura 26.** Orientación de la transcripción del gen del regulador con respecto al operón regulado. En el dibujo se muestra en gris el operón y en color el gen del regulador con las tres posibles orientaciones de la transcripción de uno con respecto al otro. En el gráfico se muestra la proporción de cada uno de los posibles sistemas en el grupo de operones que tiene el gen del regulador contiguo y en el grupo de genes que tienen el regulador separado

En el gráfico de la figura 26 se observa como la orientación de la transcripción de los operones que tienen el gen del regulador separado del operón por algún gen intermedio es aproximadamente la que se esperaría al azar, que sería 50% de los casos en los que la transcripción se produce en la misma dirección, y 25% de los reguladores transcritos en

dirección divergente al operón que regulan y otro 25% en dirección convergente. Sin embargo, cuando el gen del regulador y el operón están anexos lo más frecuente, a diferencia de lo que esperaríamos por azar, es que los genes se transcriban de forma divergente: un 66% de los casos. El 33% restante se transcribe en la misma dirección y no hay ningún caso donde la transcripción sea convergente. Estos datos se asemejan a la situación descrita en *Escherichia coli* donde el 60% se transcriben de forma convergente, el 35% en la misma dirección y un 5% de forma convergente (Warren y Wolde, 2004). Definitivamente, parece que la transcripción divergente está seleccionada en los operones que tienen el gen del regulador contiguo. Esto podría explicarse, como ya se ha sugerido anteriormente, porque está organización genética permite la co-regulación del operón catabólico y del regulador al compartirse la zona de los sitios de unión al ADN. Resumiendo lo visto, la regulación de los operones catabólicos se realiza principalmente por activadores que inducen la expresión de los promotores en respuesta a la presencia de un efector. Tanto el número de activadores como de inducciones es significativamente mayor que el que encontramos en *Escherichia coli.* Si analizamos la regulación de los promotores de operones catabólicos comparándola con la auto-regulación del gen del regulador que los regula observamos que lo más frecuente es encontrar activación del promotor catabólico frente a represión del promotor del regulador, lo que hace que el sistema sea más estable (Wall *et al.*,2004).

# Conectividad y unidades transcripcionales

Un último aspecto que nos interesa estudiar de los circuitos es su 'conectividad'. Definimos conectividad entre elementos del circuito como la cantidad de conexiones entre elementos del circuito. Las conexiones pueden ser tener dirección siendo de salida, las que parten de un nodo hacia otros y de entrada las que vienen de otros nodos hacia uno (Barabási y Oltvai, 2004). Por ejemplo, nos interesa estudiar el tamaño de los 'regulones' en biodegradación comparados con los de *Escherichia coli.* Un 'regulón' es el conjunto de genes que responde al mismo regulador y que activa o reprime un conjunto de promotores en respuesta a una señal del medio ambiente (Cases y de Lorenzo, 2005). Estudiando las conexiones de salida del regulador hacia sus genes regulados, podemos hacernos una idea del tamaño de la red transcripcional controlada por la acción de el regulador estudiado. Por otro lado, estudiando las conexiones de entrada podemos ver la cantidad de reguladores que controlan un gen y así observar como de complejo es la regulación de los genes. Al realizar el estudio de la conectividad, tanto de los reguladores de biodegradación como de los de *Escherichia coli*, no incluimos los reguladores globales, entendiendo como reguladores globales los descritos como tales para *E. coli* (Martínez-Antonio y Collado-Vides, 2003).

**Los reguladores implicados en biodegradación controlan un gran número de genes por medio de unos pocos promotores**

Para saber el tamaño de los regulones en los sistemas de biodegradación tenemos que calcular el número de genes que están controlados por un mismo regulador o, dicho de

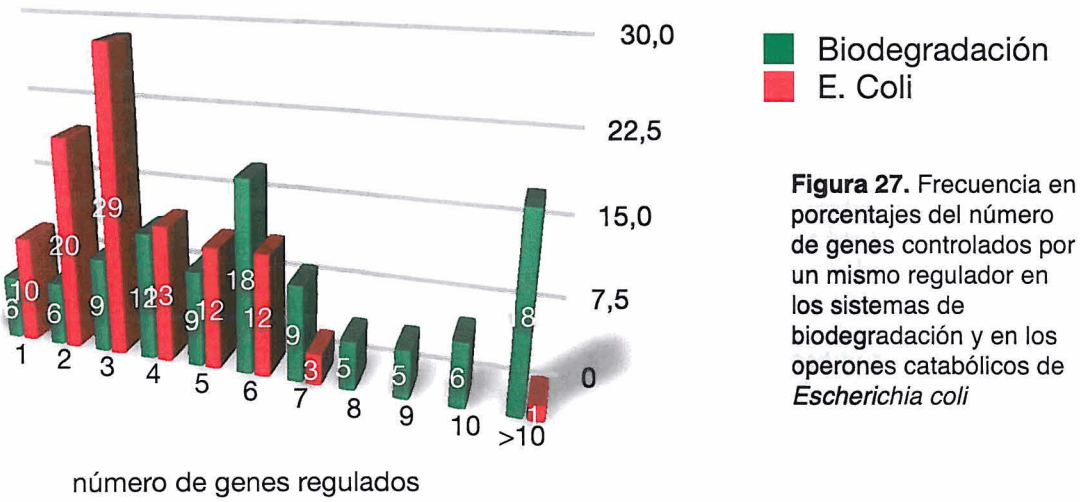



**Figura 27.** Frecuencia en porcentajes del número de genes controlados por un mismo regulador en los sistemas de biodegradación y en los operones catabólicos de *Escherichia coli*

otra manera, la conectividad entre el regulador y sus genes regulados. Comparamos los resultados con los obtenidos para *Escherichia coli.* En general, el número de genes regulados por regulador parece mayor en biodegradación (figura 27). Fijémonos en los casos extremos. Por ejemplo, en *E. coli* uno de los reguladores que controla mayor número de genes es ExuR que regula 7 genes. Este regulador controla la utilización de hexuronato. También controla 7 genes CytR que es un regulador implicado en el catabolismo de los nucleósidos y su reciclaje. En nuestra muestra de biodegradación NahR de *Pseudomonas putida* controla 23 genes.

Para comprobar si realmente los reguladores de los sistemas de biodegradación controlan un mayor número de genes hicimos el test de la 't de Student' que nos permite comparar las medias de distribuciones. Obtuvimos un valor p menor de 0,001 lo que implica que existe evidencia estadística para asegurar que el número de genes regulados por el mismo regulador es diferente en las dos muestras. La media del número de genes regulados por regulador es similar, aunque algo mayor, a la obtenida en un estudio anterior sobre la red de regulación transcripcional de *Escherichia coli* (Thieffry *et al.*,1998) a pesar de que en aquel estudio se incluían reguladores globales y operones no catabólicos: 3,8 en este frente a 3 en el estudio sobre *E. coli.* Esto es debido a que la gran mayoría de los reguladores controlan pocos genes y los valores extremos no desvían la media.

Que un regulador controle un mayor número de genes no implica que esté actuando sobre un número mayor de promotores. Para arrojar luz sobre este asunto calculamos el número de promotores por regulador, tanto en biodegradación como en *E. coli*, y los comparamos (figura 28).

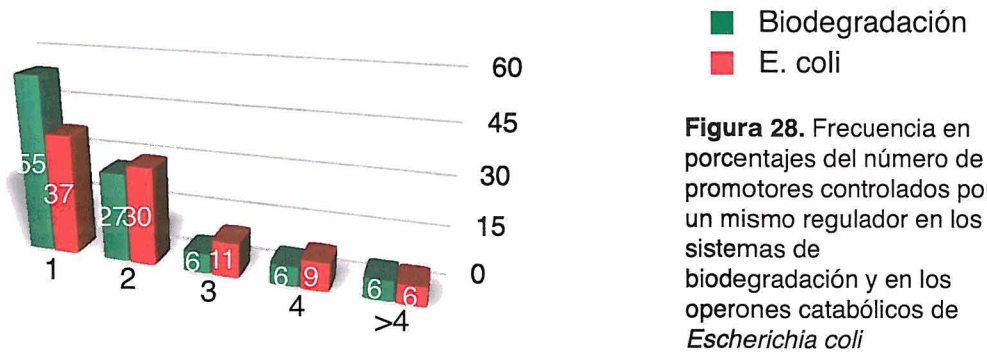



**Figura 28.** Frecuencia en porcentajes del número de promotores controlados por un mismo regulador en los sistemas de biodegradación y en los operones catabólicos de *Escherichia coli*

En la figura 28 da la impresión de que los reguladores de *Escherichia coli* son capaces de regular un mayor número de promotores. Elegimos hacer el test de Mann-Whitney para comparar las dos muestras, porque no parecen seguir una distribución normal, y comprobamos si existen diferencias significativas. Obtenemos un valor p igual a 0,03 lo que nos indica que hay suficiente evidencia estadística para afirmar que las dos muestras son diferentes. Así pues, hemos visto que los reguladores de los sistemas de biodegradación controlan un mayor número de genes a través de un menor número de promotores. Esto podría ser debido a que los operones que contienen los genes regulados sean más largos en biodegradación que en *Escherichia coli*.

**Las unidades transcripcionales son más largas en biodegradación que en *E. coli***

Pasamos a analizar los operones o unidades transcripcionales (UTs), ya que no son operones si nos atenemos a la definición clásica de operón que requiere que el operón tenga más de un gen (Jacob *et al*, 1960). Es interesante estudiar la organización en UTs de los genes pues este es el primer nivel por el que se asocian, de forma física, para transcribirse juntos. Comparamos la longitud en genes de las UTs de biodegradación frente a las de *Escherichia coli* (figura 29). Conviene comentar que estamos comparando todas las UTs, incluidas las de los genes reguladores.

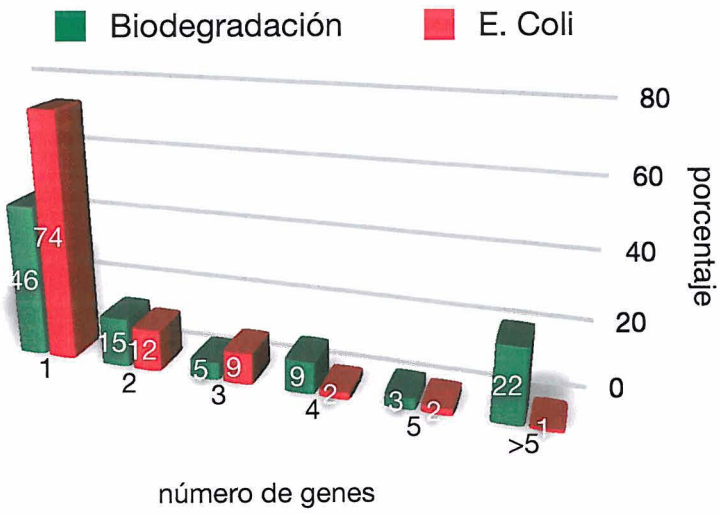



**Figura 29.** Frecuencia en porcentajes de la longitud en genes de las unidades transcripcionales de biodegradación comparadas con las de *Escherichia coli.*

En la figura da la impresión de que las UTs de biodegradación son más largas generalmente, especialmente por los 47 UTs de más de 5 genes en biodegradación frente a sólo 2 casos de esta categoría en *Escherichia coli*. Hacemos un test de Mann-Whitney comparando las longitudes de los dos grupos de UTs. Obtenemos un valor $p < 0,001$ lo

que nos permite rechazar con garantías la hipótesis de que las UTs de biodegradación sean igual de largas que las de *Escherichia coli* y afirmar que existe evidencia estadística en contra de esta afirmación. La gran cantidad de UTs de un sólo gen puede ser atribuida a la presencia de las que codifican reguladores. La mayor longitud de las UTs en biodegradación se podría atribuir a que les permite un mejor empaquetamiento de complejos y rutas catabólicas en un mismo elemento funcional sin necesidad de crear varios sitios de unión para los reguladores, ya que se ha propuesto que es más probable que se formen nuevos operones que nuevas secuencias reguladoras (Price *et al.,* 2005).
A continuación, nos pareció interesante comparar las longitudes de las UTs inducidas separándolos en dos grupos: los inducidos por medio de un activador y los inducidos por medio de un represor. La idea detrás de esto era que quizá el tener UTs muy largas reguladas por un represor podía ser una desventaja. En caso de una mutación en el sistema regulador los genes se expresarían de forma descontrolada y esto podría suponer un derroche de energía y una desventaja en una situación extrema. Hicimos un test de Mann-Whitney comparando las longitudes y obtuvimos un valor p igual a 0,84. No existe evidencia estadística para afirmar que existen diferencias entre los dos grupos de UTs.

**Los genes y promotores están controlados por pocos reguladores**

Como siguiente paso analizamos la 'conectividad' de entrada de los genes o, en otras palabras, cuantos elementos los controlan. Este es un modo de cuantificar el control ejercido sobre cada promotor y sobre cada gen calculando el número de reguladores que actúan sobre ellos (figura 30).

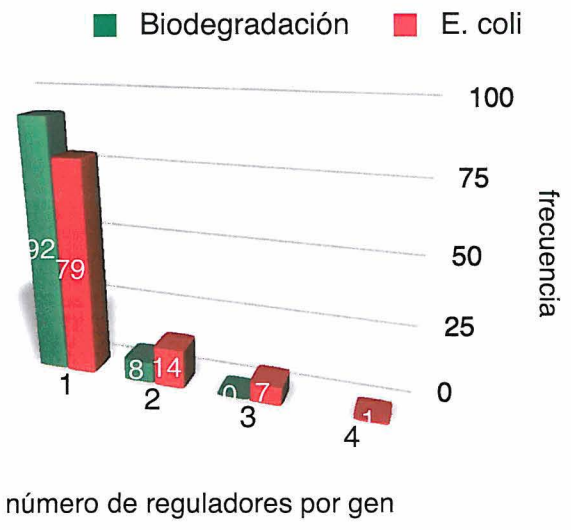



**Figura 30.** Frecuencia en porcentajes del número de reguladores que están regulando un mismo gen en los sistemas de biodegradación y en los operones catabólicos de *Escherichia coli*

Se observa que que los sistemas de biodegradación los genes son regulados en la mayoría de los casos por un único regulador. Algo similar ocurre en *Escherichia coli* pero la diferencia no parece tan rotunda. Para comprobar si hay diferencias significativas entre las dos muestras hacemos un test de Mann-Whitney. Obtenemos un valor p menor de 0,001, lo que implica que existe evidencia estadística para afirmar que las muestras son distintas. Esta diferencia no puede atribuirse a la presencia de reguladores globales ya que no los incluimos en el análisis.

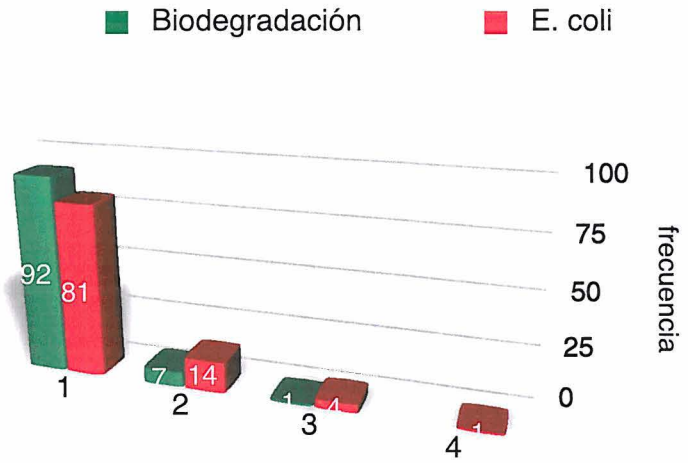



**Figura 31.** Frecuencia en porcentajes de reguladores que están regulando un mismo promotor en los sistemas de biodegradación y en los operones catabólicos de *Escherichia coli*

En el caso de los promotores parece ocurrir algo similar a lo que ocurre con los genes (figura 31): la mayoría de lo promotores tanto en *E. coli* como en biodegradación, están controlados por un único regulador pero hay más casos de promotores controlados por más de un regulador en *E. coli.* Calculamos un test de Mann-Whitney para ver si la diferencia entre las muestras es significativa. Obtuvimos un valor p igual a 0,01 lo que nos indica que existe evidencia estadística suficiente para afirmar que las dos muestras son diferentes. Por lo tanto, podemos afirmar que los promotores de biodegradación están controlados por un menor número de reguladores que los de *Escherichia coli.* Concluyendo, los genes y promotores en biodegradación tienen un control transcripcional más sencillo que en *Escherichia coli* y que consiste principalmente en estar agrupados en UTs largas.

**Los genes se expresan desde menos promotores distintos en biodegradación**

Expresar un gen desde varios promotores distintos puede servir para añadir un nivel adicional de control regulatorio modulando el nivel de expresión de cada uno de estos promotores. Por ello, para terminar con esta parte del análisis de los promotores queremos observar el número de promotores desde los que se expresa un gen y, de nuevo, compararlo con lo que se da en *Escherichia coli* (figura 32).

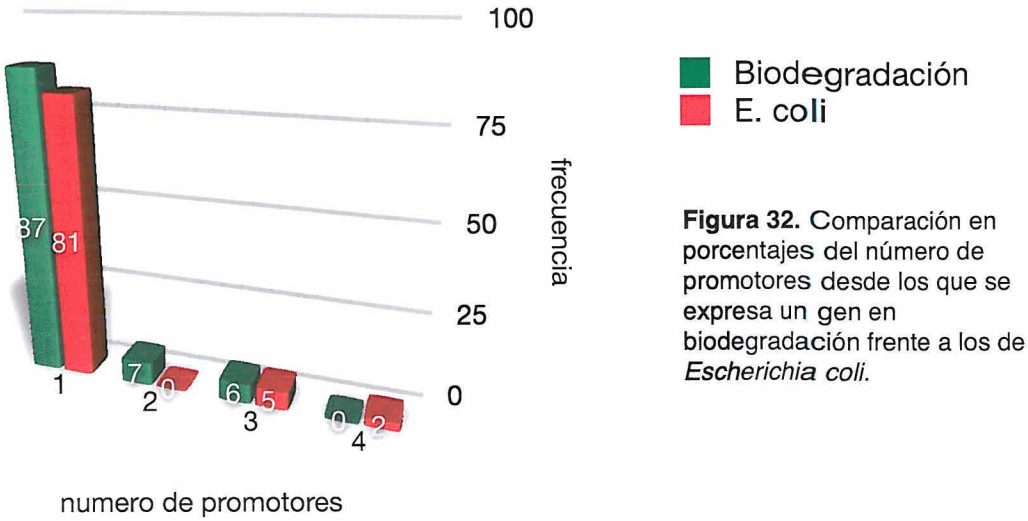



**Figura 32.** Comparación en porcentajes del número de promotores desde los que se expresa un gen en biodegradación frente a los de *Escherichia coli*.

No se observan grandes diferencias ya que en ambos casos la gran mayoría de los genes se expresan desde un sólo promotor. Hacemos el test de Mann-Whitney que no requiere condiciones para su aplicación para comparar los dos grupos porque los datos no siguen una distribución normal. Obtenemos un valor p = 0,023 lo que nos indica que hay diferencias significativas entre los dos grupos. Los genes de *Escherichia coli* no se expresan desde la misma cantidad de promotores que los de biodegradación y los datos parecen apuntar a que los genes de biodegradación tienden a expresarse desde un menor número de promotores. Así también a este nivel parece que en biodegradación el control de la transcripción es mas sencillo.

En resumen, podemos destacar de la caracterización de la conectividad de los circuitos de biodegradación varios puntos. Los regulones comprenden muchos más genes que los de *Escherichia coli* por el mayor tamaño de los operones. En cuanto a la conectividad de los reguladores y de los promotores y genes regulados, observamos que en los sistemas de biodegradación los reguladores controlan un mayor número de genes utilizando un menor número de promotores y que tanto promotores como genes están controlados por

un mayor número de reguladores en *E. coli*. Todo unido parece indicar que la regulación en *E. coli* es más compleja mientras que en biodegradación los sistemas son más estables y compactos, existen menos requerimientos para el control de los promotores y un único estímulo provoca una respuesta transcripcional mucho mayor.

# Integración con la fisiología

### En biodegradación hay una mayor proporción de promotores asociados a sigma 54

Otro aspecto interesante del estudio de los promotores es su asociación con los factores sigma. Los factores sigma son factores de iniciación de la transcripción que controlan la unión de la ARN-polimerasa al promotor. En respuesta a las diferentes condiciones ambientales se activan diferentes factores sigma, por ejemplo sigma 32 en respuesta al choque térmico o sigma 38 en la fase estacionaria o en caso de inanición (Paget y Helmann, 2003). Sigma 70 es el que transcribe la gran mayoría de los genes y sigma-54 es un caso especial por no ser homólogo a los demás, tener un mecanismo de acción distinto y ser más flexible en cuanto a sus funciones reguladas (Reitzer y Schneider, 2001). Comparamos la proporción de los distintos tipos de sigma en biodegradación frente a lo que encontramos para *Escherichia coli* (figura 33).

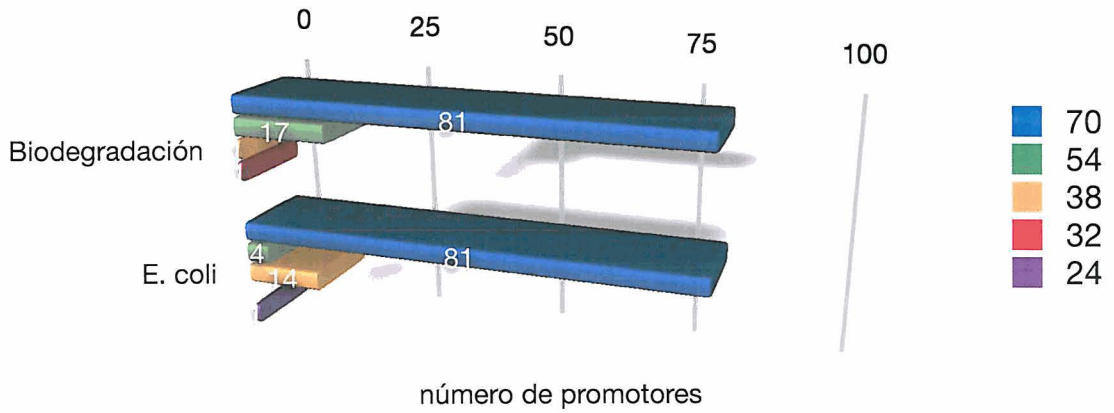



**Figura 33.** Comparación en porcentajes de los tipos de factores sigma asociados a lo promotores en biodegradación frente a los de *Escherichia coli*

Observamos que la principal diferencia es la mayor proporción de factores sigma 54 en los promotores de biodegradación. Para ver si estas diferencias de proporción son significativas hacemos un test ji-cuadrado. El test nos da un valor p = 0,003 lo que nos indica que hay diferencias estadísticas significativas entre las dos muestras y que los promotores dependientes de sigma 54 están más representados en biodegradación que en *Escherichia coli*. Una razón para explicar esto puede estar en el hecho de que los grupos de genes de biodegradación están asociados frecuentemente a elementos móviles que pasan de unas bacterias a otras. El factor sigma 54 frecuentemente se ha visto

asociado la integración con la fisiología del huésped (Cases y de Lorenzo, 1998; Carmona *et al.*, 2000; Solera *et al.*, 2004) y quizá por su ventaja adaptativa a la hora de integrar las nuevas funciones con el metabolismo del huésped este tipo de promotores pueden haber sido seleccionados a favor. Aunque también podría ser que estuviéramos ante un efecto fundador: un operón regulado por el factor sigma 54 habría sido el tipo de promotor utilizado originalmente para expresar un conjunto de genes y su uso se habría extendido, no por ofrecer una ventaja adaptativa, si no por ser el único disponible.

**La mayor proporción de promotores asociados a sigma 54 en biodegradación no está causada por un efecto fundador**

Para comprobar si la mayor proporción de promotores asociados a sigma 54 pudiera estar originada por un efecto fundador comprobamos la homología entre los genes de las UTs que tienen un promotor de este tipo asociado. Para considerar dos UTs homologas al menos dos tercios de sus genes deben de tener un 30% o mas de identidad de secuencia (ver detalles en Métodos) ya que con un 30% de identidad se mantiene la estructura de las proteínas y se considera que tienen un origen evolutivo común (Devos y Valencia, 2000). Sólo identificamos pequeños grupos de TUs homólogas reguladas por sigma 54 por lo que parece que se puede descartar un efecto fundador. Esto indica que existe una convergencia evolutiva en la utilización de sigma 54 para la expresión de los promotores de estas TUs, de lo que se deduce que confiere alguna ventaja que ha sido seleccionada. Quizá una explicación sea, como es mencionado anteriormente, su relación con la integración con la fisiología del huésped.

# Movilidad y organización genética

Desde que apareciera el trabajo pionero de Jacob & Monod hasta el día de hoy se ha venido aceptando el hecho de que la agrupación de genes en operones es la principal forma tanto de organización genética como de co-regulación (Jacob *et al.*, 1960). Sin embargo, análisis más recientes permiten observar la organización del genoma con mucho mayor detalle. Algunos de estos estudios afirman que los genes que están relacionados funcionalmente tienden a estar agrupados en el ADN (Salgado *et al.*, 2000; De Daruvar *et al.*, 2002). Otros estudios afirman que, en *Escherichia coli*, los operones que regulan unos a otros y los operones co-regulados están más cerca entre sí en el genoma de lo que sería de esperar por el azar (Warren y Wolde, 2004). También sabemos que los genes del regulador y sus operones regulados están más cerca entre sí si actúan en respuesta a estímulos externos y más alejados si actúan ante estímulos internos (Janga *et al.*, 2007). Por otro lado, los genes implicados en biodegradación se encuentran frecuentemente incluidos en elementos móviles (Springael y Top, 2004) y se estima que estos elementos móviles juegan un papel fundamental en la formación de nuevas rutas que se ensamblan por un proceso que implica transferencia horizontal de genes(HGT) (Top y Springael 2003; Van der Meer y Sentchilo, 2003). Nos preguntamos si la organización genética de los genes, su orden en el genoma, puede haber influido de alguna manera favoreciendo estos procesos. Por ello estudiamos la posición de los genes catabólicos en los operones. Una aclaración: aquí sustituimos el concepto introducido anteriormente de unidad transcripcional por el de operón ya que en esta ocasión si estamos hablando de operones en el sentido 'clásico' del término.

### Los complejos enzimáticos suelen estar codificados en un único operón

Anteriormente observábamos que el regulador se encontraba frecuentemente contiguo, o en su defecto muy cerca, de los genes regulados. Esto permite tener una estructura compacta en la que la regulación y las enzimas están contenidos en el mínimo espacio posible. Pero no sabemos si realmente todos los genes que codifican para una enzima formada por varias proteínas están contenidos en el mismo operón así que lo investigamos. 55 de las 57 enzimas compuestas por más de una proteína tienen los genes que codifican para esas proteínas contenidos en el mismo operón, lo que supone un 96% del total de las enzimas. Uno de los casos en que los genes están distribuidos es

más de un operón es el del complejo formado por los genes *paaG, paaH, paaI, paaJ y paaK* en *Pseudomonas putida* U. Este complejo degrada fenilacetato y tiene distribuidos los genes entre los operones *paaFGHI* y *paaJKL*. Los genes que codifican para el complejo están situados de forma continua en el ADN pero están agrupados en estos dos operones. Todos los genes son subunidades del complejo que oxida el anillo del fenilacetato (Olivera *et al.*, 1998) . El otro complejo que tiene sus genes repartidos en dos operones es el formado por las proteínas TutD, TutE, TutF y TutG y degrada tolueno en *Thaeura aromatica* T1. TutE es homólogo a enzimas que activan la piruvato formato-liasa y TutD es homólogo a enzímas con actividad piruvato formato-liasa. Los genes en este caso están distribuidos entre los operones *tutE y tutFDGH.* En este caso también están situados de forma continua en el ADN pero existe un promotor que expresa *tutFDGH* detrás de *tutE* (Coschigano *et al.*, 1998).

**Los genes que codifican un complejo enzimático suelen estar agrupados unos junto a otros dentro de los operones**

Ya sabemos que en la mayoría de los casos las enzimas agrupan los genes que las codifican en un mismo operón pero no sabemos si dentro del operón los genes se colocan de forma ordenada. ¿Están localizados unos junto a otros los genes que codifican para el mismo complejo? Se ha observado que los operones tienen una tendencia a contener genes relacionados funcionalmente (Rocha, 2006). Comprobamos si en lo operones catabólicos de biodegradación es así. Para hacer este análisis calculamos la cantidad de genes que forman parte de un complejo compuesto por varias proteínas están situados en los operones al lado de otro gen que codifique para el mismo complejo (figura 34).

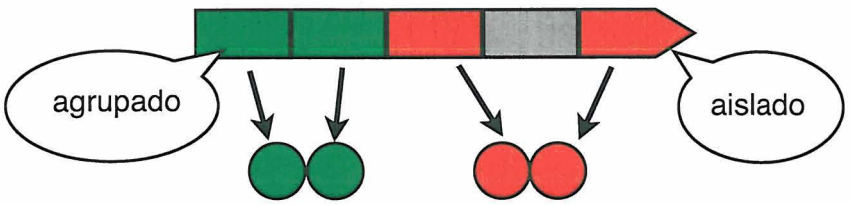



**Figura 34**. En la figura se muestra un operón que codifica dos complejos: complejo verde y complejo rojo. Los genes que codifican para el complejo verde tienen al lado un gen que codifica para el mismo complejo que ellos y los clasificamos como agrupados. Los genes que codifican el complejo rojo no tiene ningún gen que codifique para el mismo complejo que ellos a su lado así que los clasificaríamos como aislados

Si un gen tiene a su lado otro gen que codifique para el mismo complejo que él lo clasificamos como agrupado, si no como aislado. De esta manera obtenemos que 232 de

un total de 238 genes que codifican para complejos enzimáticos de varias proteínas están agrupados, lo que supone un 97,5% de los genes. Tan sólo 6 genes están aislados. Este resultado podría estar causado por otros factores. Por ejemplo que los operones contuvieran únicamente los genes de un único complejo: en esta situación el orden no importaría ya que todos los genes de cada complejo siempre estarían al lado de algún gen que también perteneciera al complejo. Para comprobar que este tipo de situaciones no estuviera sesgando nuestros resultados comparamos la proporción de genes agrupados en los operones frente a la proporción de genes agrupados en los mismos operones con sus genes re-ordenados al azar. Re-ordenamos los genes en los operones y calculamos la proporción agrupados probando 1000 ordenes distintos. Comparando la proporción real frente a la distribución de proporciones generadas al hacer obtenemos un valor z de 343. Al convertir este valor z en un valor p obtenemos un valor p menor de 0,001 por lo que podemos asegurar que existe evidencia estadística para afirmar que con el orden real los genes están más agrupados que ordenados al azar.

**Los genes que codifican complejos enzimáticos que realizan reacciones consecutivas suelen estar ordenados de forma consecutiva en los operones**

Ya hemos visto que los genes que codifican para un mismo complejo enzimático suelen estar agrupados dentro de un mismo operón. Pero, ¿están ordenados según las reacciones que realizan los complejos enzimáticos que codifican? La figura 35 (siguiente página) explica esta idea. Como se puede ver en la figura, si un complejo genera un producto que es un sustrato del siguiente complejo codificado en el operón en el sentido de la transcripción definimos la conexión entre ellos como directa. Sin el complejo que genera el producto está detrás del anterior en el sentido de la transcripción la conexión es inversa. Si no es ninguno de estos casos lo clasificamos como sin conexión. Analizando de esta manera los operones encontramos una proporción de un 72% de conexiones totales (conexiones/conexiones posibles), siendo un 62% de conexiones directas (conexiones directas/conexiones posibles) y un 10% de conexiones indirectas (conexiones inversas/conexiones posibles).

Parece que los complejos están bastante conectados pero para comprobar si están más o menos conectados que lo que sería de esperar al azar calculamos de nuevo las conexiones re-ordenando los genes. Re-ordenamos los genes al azar 1000 veces y para cada una de las veces calculamos la proporción de conexiones totales, directas e inversas. Posteriormente hicimos la distribución de cada una de ellas y calculamos el

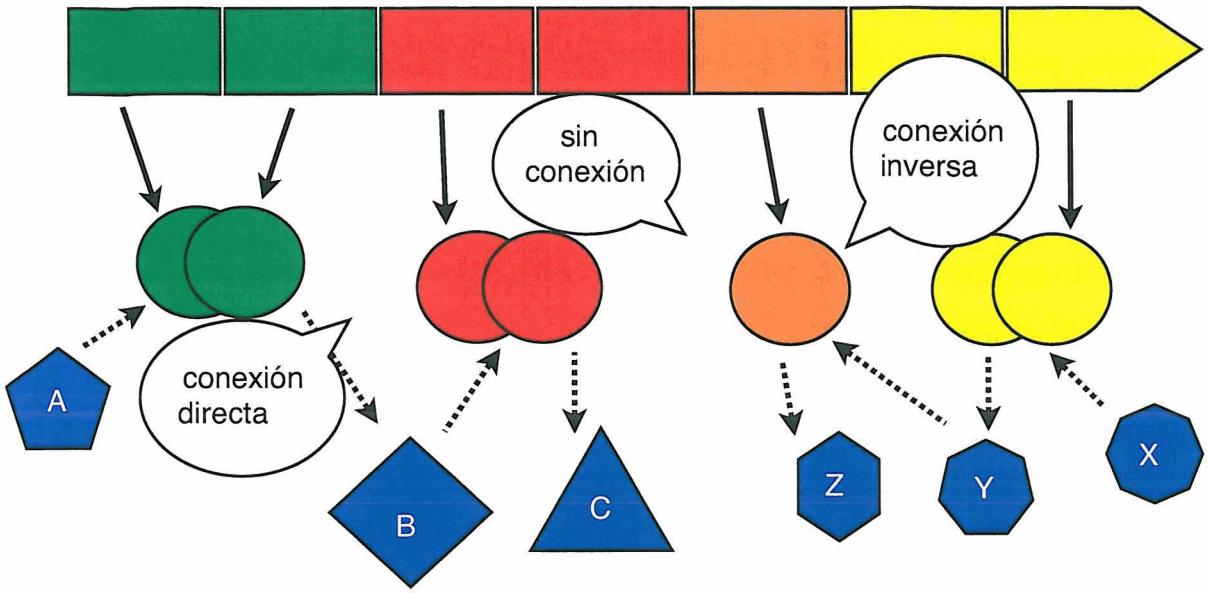



**Figura 35.** Posibles conexiones entre los complejos enzimáticos codificados por los operones. Siguiendo la dirección de la transcripción, si un complejo (complejo verde) genera un producto (rombo B) que es sustrato del siguiente complejo (complejo rojo) los dos complejos tiene una conexión directa. Si un complejo (complejo amarillo) genera un producto ( heptágono Y) que es sustrato del complejo anterior según el sentido de la transcripción (complejo naranja) los dos complejos tienen una conexión inversa. Si un complejo (complejo rojo) genera un producto (triángulo C) que no es sustrato del siguiente complejo los dos complejos no tienen conexión. En este ejemplo habría 3 posibles conexiones entre los complejos, 1 'conexión directa', 1 'conexión inversa' y una 'sin conexión'. Las proporción de conexiones totales sería 2/3 , de directas 1/3 y de inversas 1/3.

valor z de las conexiones reales frente a las distribuciones generadas re-ordenando los genes. Para las conexiones totales obtuvimos un valor z de 285. Esto nos permite afirmar que el orden real permite realizar un mayor número de conexiones entre complejos contiguos que las que se formarían al azar. Para las conexiones directas obtuvimos un valor z de 365. Podemos también afirmar, por lo tanto, que el orden real permite realizar más conexiones directas que las que se podrían formar al azar. Finalmente, para las conexiones inversas obtuvimos un valor z de -96,5. Este valor nos indica que este tipo de conexiones aparece con menos frecuencia con el orden real de lo que se esperaría que ocurriera al azar.

En resumen, los complejos enzimáticos están organizados en grupos de genes en los operones que están ordenados en el sentido de la transcripción. Este orden permite que los productos generados por la acción de un complejo sean el sustrato del siguiente complejo codificado por el operón. Los complejos enzimáticos que están conectados en el orden contrario a la transcripción aparecen menos que lo que sería de esperar al azar. Da la impresión de que los operones que contienen complejos completos y conectados entre sí en la dirección de la transcripción han sido seleccionados a favor posiblemente por ser más eficientes en la expresión de las rutas de degradación de los compuestos.

**El orden de los genes en los operones permite transferir tanto un mayor número de complejos enzimáticos como de complejos que realizan reacciones consecutivas**

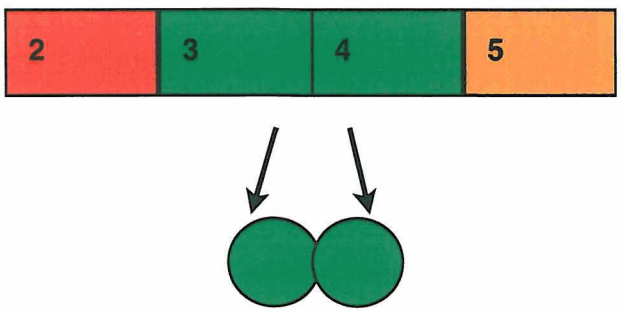
La transferencia horizontal de genes (HGT) es frecuente entre especies bacterianas (Springael y Top,2004;Top y Springael, 2003) y está causada por la capacidad de las bacterias de transferir ADN por medio de procesos de conjugación e integrar ADN externo a través de un proceso llamado transformación. Esta movilidad del ADN podría ser la explicación de la presencia de conjuntos de genes muy parecidos en bacterias muy alejadas entre sí desde el punto de vista evolutivo (Furukawa *et al.*,2004). Es un hecho conocido que los operones catabólicos de biodegradación se encuentran frecuentemente contenidos en elementos móviles como plásmidos o transposones (Van der Meer y Sentchilo, 2003). Por nuestra parte, hemos observado que los genes que codifican para los complejos enzimáticos siguen un orden concreto dentro de los operones. Pensando en todos estos procesos de transferencia de ADN entre bacterias nos planteamos si este orden de los genes favorecería la adquisición de complejos enzimáticos completos por medio de eventos de HGT. Para responder a esta pregunta diseñamos un experimento de simulación de un evento de HGT (figura 36).

**1** Coge cada operón. **Por ejemplo, este**


**2** Selecciona un inicio para la transferencia en función de la longitud del operón (en este caso entre -6 y 6). **Por ejemplo, 2**

**3** Selecciona una longitud para la transferencia siguiendo una distribución elegida (normal, exponencial o uniforme): **Por ejemplo, 4**

**4** Selecciona el fragmento a transferir que empieza en ese inicio y tiene esa longitud en genes.




**5** Cuenta el número de complejos completos transferidos. **En el ejemplo, 1**

**Figura 36.** Simulación de HGT. Para simular un evento de transferencia horizontal de genes seguimos los siguientes pasos. 1, cogemos un operón catabólico. 2, cogemos un fragmento del operón. 3, comprobamos el número de complejos transferidos.

La idea es coger un operón catabólico y seleccionar un conjunto de genes del operón que representarán el ADN adquirido por HGT. Empezamos eligiendo un "principio de transferencia" al azar que será un gen situado entre un número de genes igual a la longitud en genes del operón antes del primer gen del operón y el último gen del operón. Escogimos esta forma de seleccionar el inicio de transcripción porque nos pareció más real que el inicio del fragmento de ADN adquirido no tuviera porque estar dentro del operón. A continuación elegimos una longitud determinada en genes para el fragmento de ADN a transferir. Clasificamos las simulaciones en tres tipos dependiendo del sistema elegido para seleccionar los fragmentos de ADN a transferir. Una de las posibilidades es obtener la longitud de los fragmentos de ADN transferidos en función de una distribución normal. De esta forma, cada vez que seleccionamos la longitud de un fragmento de ADN a transferir esta se genera por un número al azar obtenido de una distribución normal con media 4, que se aproxima a la longitud media de los operones catabólicos de biodegradación, y desviación estándar 1.4, que nos permite obtener valores variados pero no negativos. De forma similar hicimos otras dos simulaciones. En el segundo caso obtuvimos la longitud de los fragmentos de ADN a partir de una distribución exponencial de nuevo de media 4 y desviación estándar 1.4. En el tercer caso se obtenían a través de una distribución uniforme seleccionando un número al azar del 1 al 15, que es el número de genes del operón más largo en biodegradación (ver detalles en Métodos). Hicimos este proceso 1000 veces para cada operón catabólico y con los tres distribuciones distintas para seleccionar la longitud de los fragmentos de ADN transferidos. Para determinar si la cantidad de complejos completos transferidos era diferente a la que se habría obtenido con un orden de genes al azar hicimos este mismo experimento probando a re-ordenar los operones y repitiendo el experimento probando 1000 ordenes distintos para cada operón.

En cada simulación de un evento de transferencia obtenemos un número de complejos transferidos que podemos dividir por el número de posibles complejos a transferir para cuantificar lo productivo que ha sido el evento. Así obtendremos un número entre cero y uno, cero si no han pasado complejos, uno si han pasado todos los que podían pasar. De este modo tras realizar 1000 simulaciones para un operón podemos calcular la media de la 'transferibilidad' de ese operón. Finalmente, si comparamos esa media con las medias obtenidas para los 1000 ordenes alternativos probados para el operón podemos calcular un valor z. Este valor z cuantifica si el orden real de los operones permite 'transferir' un mayor número de complejos. En la figura 37 se representan las proporciones de valores z obtenidos utilizando las distintas distribuciones para coger los fragmentos de ADN.

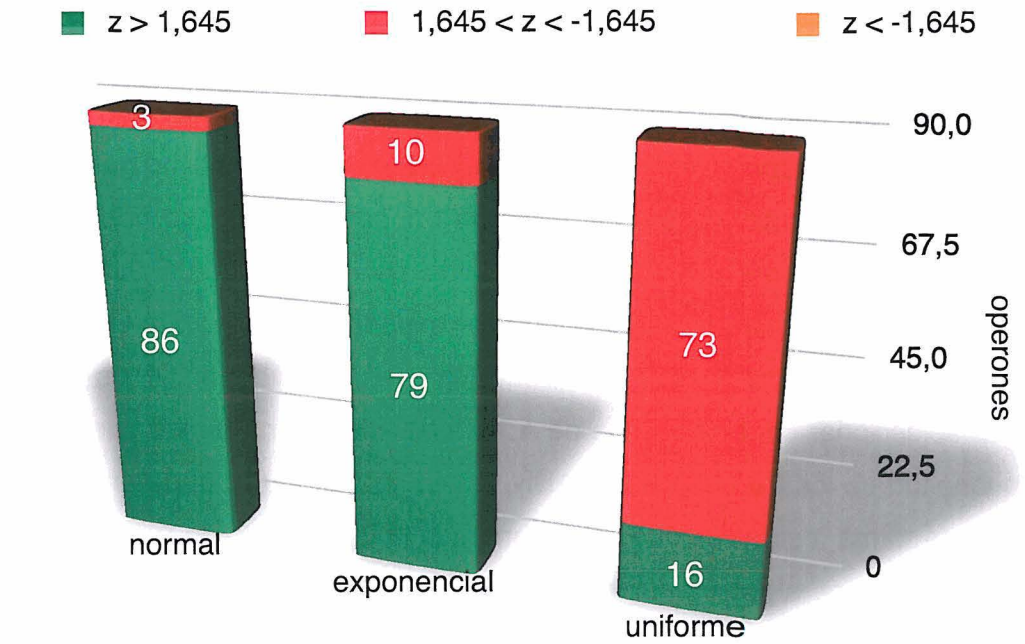



**Figura 37.** Clasificación en categorías de los operones basadas en los valores z obtenidos al comparar la proporción de complejos transferidos con el orden real del operón frente a 1000 ordenes aleatorios. Si el orden real del operón es significativamente mejor que los ordenes al azar, $z > 1,645$, está representado en color verde, si el orden no es significativamente ni mejor ni peor está representado en rojo y si es significativamente peor que el orden al azar al transferir complejos, $z < -1,645$, está representado en naranja. Las tres columnas representan el tipo de distribución utilizada para seleccionar el fragmento de ADN con el que se simula la transferencia

En la figura podemos ver que en ningún caso el orden del operón es peor que el orden al azar transfiriendo complejos. Tanto si cogemos los fragmentos de ADN siguiendo una distribución normal como una exponencial, la gran mayoría de operones transfieren un mayor número de complejos con el orden real que con un orden al azar. Utilizando la distribución normal el 97% de los operones transfiere mejor complejos completos con el orden real y utilizando la exponencial un 89%. Curiosamente, si utilizamos una distribución uniforme para seleccionar la longitud de los fragmentos de ADN a transferir el orden de los genes no parece tan importante ya que tan sólo un 18% de los operones transfiere mayor cantidad de complejos de forma significativa que los mismos operones con sus genes re-ordenados.

En la sección anterior observamos como los genes que codifican para complejos enzimáticos suelen estar agrupados en un mismo operón y ordenados de forma que los complejos que realizan reacciones consecutivas suelen estar ordenados en el operón en la dirección de la transcripción. Esto sugiere que quizá también la transferencia de complejos conectados sea mayor con el orden real que con otro orden aleatorio. Para

comprobarlo realizamos una simulación similar a la anterior en la que al final contábamos la proporción de complejos conectados entre sí por ser un producto de uno de ellos sustrato del otro. Los resultados se muestran en la figura 38.

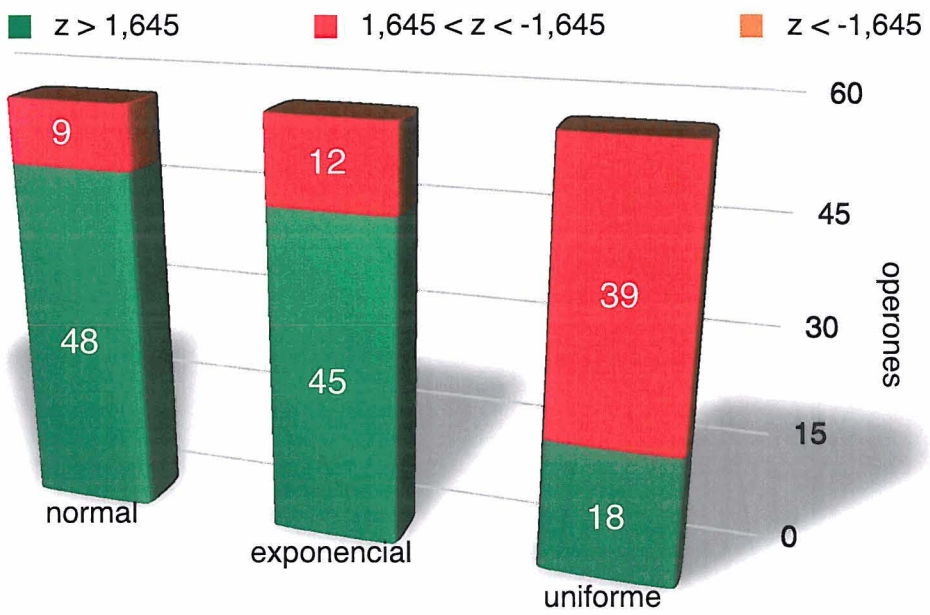



**Figura 38.** Clasificación en categorías de los operones basadas en los valores z obtenidos al comparar la proporción de complejos transferidos conectados entre sí con el orden real del operón frente a 1000 ordenes aleatorios. Si el orden real del operón es significativamente mejor que los ordenes al azar, z > 1,645, está representado en color verde, si el orden no es significativamente ni mejor ni peor está representado en rojo y si es significativamente peor que el orden al azar al transferir complejos, z < -1,645, está representado en naranja. Las tres columnas representan el tipo de distribución utilizada para seleccionar el fragmento de ADN con el que se simula la transferencia

De nuevo no existe ningún caso en el que el orden del operón sea peor que un orden aleatorio transfiriendo complejos conectados entre sí. También en este caso si cogemos la longitud de los fragmentos de ADN siguiendo una distribuciones normal o una exponencial obtenemos que pasan más complejos conectados entre sí en la mayoría de los casos. Usando la distribución normal obtenemos que un 84% de los operones transfieren más complejos conectados de forma significativa con el orden real. Usando la distribución exponencial el porcentaje se reduce hasta el 79%. Aquí también si utilizamos una distribución uniforme para coger los fragmentos de ADN no encontramos tantas diferencias entre el orden real y los aleatorios y tan sólo el 31% de los operones transfieren más complejos conectados de forma significativa. Tanto en este caso como en

el de la transferencia de complejos enzimáticos completos, esto podría ser un efecto causado por la propia naturaleza de la distribución uniforme. En esta distribución hay una probabilidad mucho mayor que en la normal o en la exponencial de transferir operones más largos o incluso completos. Si transferimos el operón completo el orden de los genes deja de ser importante porque todos están disponibles (figura 39).

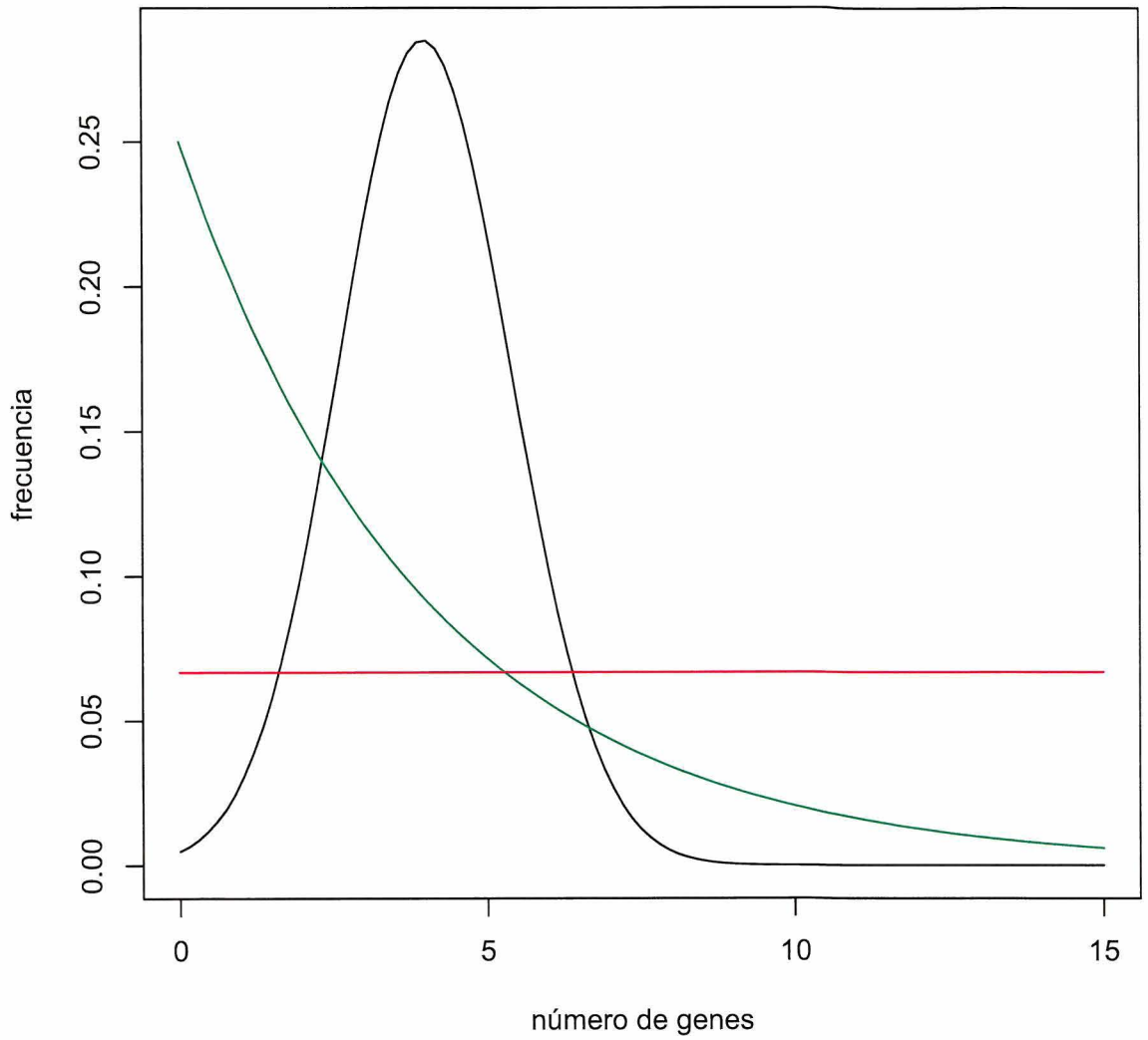



**Figura 39.** Comparación de los tres tipos de distribuciones utilizadas para escoger los fragmentos de DNA que van a ser transferidos en la simulación. La distribución normal está representada por la línea negra, la exponencial por la verde y la uniforme por la roja. Como se observa en el gráfico, con la distribución uniforme existen muchas más probabilidades de seleccionar fragmentos de ADN más largos. Al escoger fragmentos más largos hay más posibilidades de transferir operones completos. En caso de transferir operones completos el orden de los genes no influiría en la cantidad de complejos completos o conectados transferidos al estar todo los genes disponibles.

Resumiendo, estudiando la organización genética hemos encontrado que los operones catabólicos que realizan procesos de biodegradación tienen frecuentemente el gen del

regulador situado de forma contigua al operón y transcrito divergentemente. Esto facilita la co-regulación del gen de regulador y el operón regulado y agrupa en el mínimo espacio el circuito regulatorio junto con las genes regulados. Estudiando estos genes regulados hemos visto que todos los genes que codifican para un mismo complejo enzimático suelen estar codificados en el mismo operón y agrupados unos junto a otros. Estos grupos de genes tienden a estar ordenados en la dirección de la transcripción de forma que se expresan en orden complejos enzimáticos que realizan reacciones consecutivas. Si se transfieren fragmentos de estos operones al azar, este orden favorece que se transfieran más complejos completos y conectados entre sí. Se podría sugerir que la frecuente presencia de estos genes en elementos móviles ha hecho que se seleccionara un orden en los genes y una organización genética compacta que transmite más funciones utilizables por el organismo que recibe el ADN con fragmentos más cortos de ADN. En menos espacio hay más complejos completos, más complejos conectados e incluso un circuito de regulación coordinada.

# Respuesta a nuevas señales, especificidad y coevolución entre regulación y metabolismo

El vertido masivo de compuestos aromáticos en el medio ambiente producidos por la actividad humana es un hecho relativamente reciente, ya que ha ha venido ocurriendo desde la revolución industrial hasta el día de hoy. Es por eso que los microorganismos que son capaces de degradar este tipo de compuestos han tenido que integrar nuevas señales aparecidas en su medio ambiente para responder a la presencia de estos nuevos compuestos. La detección de estas señales por un regulador transcripcional puede desencadenar la expresión de un promotor que exprese unos genes que codifiquen las enzimas que degraden los compuestos que han generado la señal. Pero, ¿cómo hacen estos microorganismos para responder a estas nuevas señales? Se ha propuesto que la capacidad de respuesta ante nuevas señales medioambientales se podría desarrollar por la escasa especificidad de los reguladores y promotores preexistentes. Cuando es necesario se van haciendo más específicos suprimiendo las señales que no son deseables y ajustando la respuesta a compuestos químicos de interés (de Lorenzo y Pérez-Martín,1996). Para contrastar esta hipótesis hemos intentando cuantificar la especificidad de los sistemas de regulación, especialmente de los reguladores.

### Tanto los efectores como los reguladores son más promiscuos en biodegradación

Como primera aproximación al análisis de la especificidad de los reguladores medimos la interacción entre reguladores y efectores en nuestra muestra de reguladores de procesos de biodegradación. A continuación, comparamos los resultados con los obtenidos calculando esta misma interacción en los reguladores de los operones catabólicos de *Escherichia coli*. Queremos ver si efectivamente los reguladores de los sistemas de biodegradación son más promiscuos (figura 40). En *Escherichia coli* sólo existen 3 casos, de un total de 48 analizados, en los que un regulador interacciona con más de un efector, con dos en concreto. Sin embargo, en los sistemas de biodegradación un 37% de los reguladores interactúan con más de una molécula efectora. Un 12%, 8 casos, interactúa con más de 4 efectores diferentes. De hecho existe un caso en el que un regulador interactúa con 10 inductores distintos: NtdR de *Acidovorax sp.* JS42. Para comprobar si existen diferencias estadísticas entre los dos grupos aplicamos el test de la Mann-Whitney

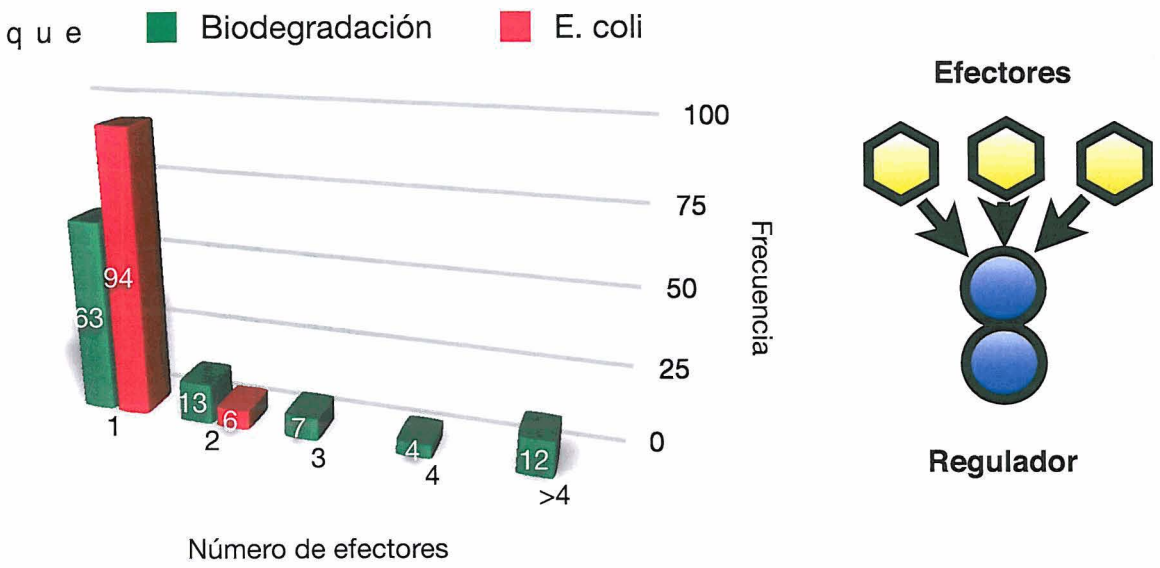



**Figura 40.** Proporción en porcentajes del número de efectores que interactúan con un mismo regulador. En la gráfica se comparan la cantidad de efectores que interactúan con un regulador en biodegradación con lo que ocurre en los reguladores que controlan operones catabólicos de *Escherichia coli*. Cada categoría del eje de abscisas representa el número de efectores con los que interactúa el regulador asignado a esa categoría. El eje de las ordenadas representa la cantidad de reguladores en esa categoría: por ejemplo, hay 8 reguladores en biodegradación que interactúan con más de 4 efectores.

no requiere condiciones para su aplicación ya que los datos no siguen una distribución normal. Existen diferencias significativas entre los dos grupos ($p < 0,001$) por lo que podemos afirmar que los reguladores que actúan sobre sistemas de biodegradación son mucho más promiscuos que sus equivalentes en *Escherichia coli*.

Observemos el punto de vista inverso: número de reguladores con los que interactúa un mismo compuesto efector. Esta perspectiva nos permite encontrar los compuestos más promiscuos. La capacidad de estos compuestos de activar un gran número de rutas catabólicas justifica el interés por identificarlos. Los genes que se expresan en respuesta a estos estímulos forman un 'estimulón ' y pueden estar controlados por diversos reguladores(Cases y de Lorenzo, 2005). La comparación entre la promiscuidad de los compuestos efectores en biodegradación y en *Escherichia coli* nos da una idea de la diferencia de magnitud de los 'estimulones' entre los dos sistemas (figura 41). Los datos obtenidos apuntan a que, en los sistemas de biodegradación, los efectores interactúan con un mayor número de reguladores distintos formando 'estimulones' de mayor tamaño que en *Escherichia coli.* En los sistemas de biodegradación existen cinco casos de efectores que interactúan con más de 4 reguladores distintos. Para comprobar si existen diferencias estadísticas entre los dos grupos aplicamos de nuevo el test de Mann-Whitney ya que en este caso los datos tampoco seguían una distribución normal. Obtuvimos un

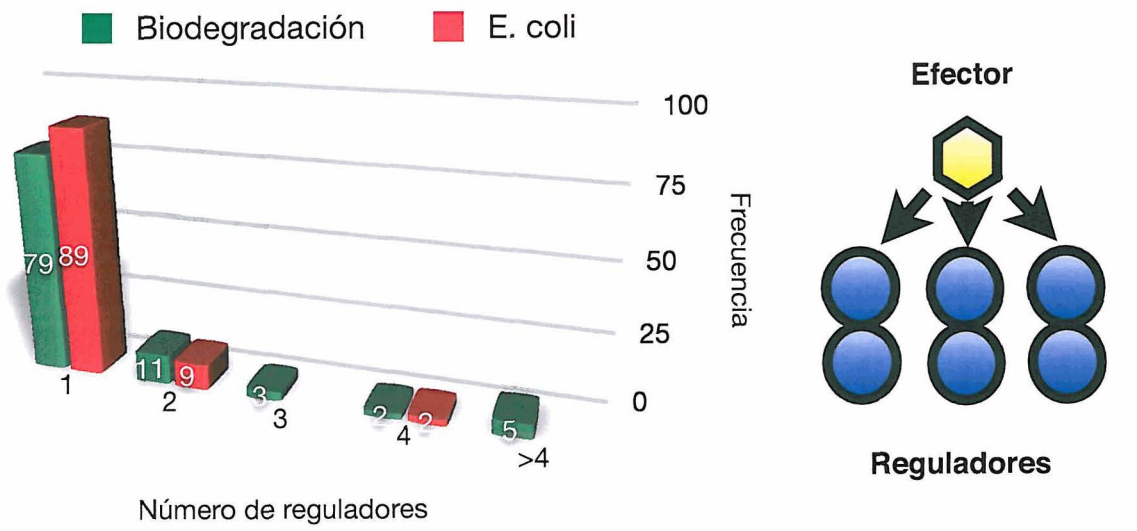



**Figura 41.** Proporción en porcentajes del número de reguladores que interactúan con un mismo efector. En la gráfica se comparan la cantidad de reguladores que interactúan con un mismo efector en biodegradación con lo que ocurre en los efectores que inducen operones catabólicos de *Escherichia coli*. Cada categoría del eje de abscisas representa el número de reguladores con los que interactúa el efector asignado a esa categoría. El eje de las ordenadas representa la cantidad de efectores en esa categoría: por ejemplo, hay 5 efectores en biodegradación que interactúan con más de 4 reguladores.

valor p = 0,13. Esto significa que no se puede rechazar la hipótesis de que las dos muestras procedan de la misma población, es decir, no encontramos diferencias significativas. Podemos concluir, por lo tanto, que la aparición de moléculas que interaccionan con los reguladores no parece afectar a un mayor número de reguladores en los sistemas de biodegradación que en *Escherichia coli*, por lo tanto no podemos descartar que los 'estimulones' sean similares en ambos casos.

### Un mismo regulador es capaz de responder a muchos compuestos diferentes

Hemos observado que los reguladores implicados en biodegradación son capaces de interaccionar con un número significativamente mayor de efectores que los que controlan el catabolismo en *Escherichia coli*. Pero, ¿cómo son de diferentes entre sí estos efectores? Nos gustaría, de algún modo, cuantificar la especificidad de los reguladores. La manera que encontramos de cuantificar la capacidad de responder a diferentes señales de los reguladores fue calculando la diferencia entre los compuestos químicos a los que es capaz de responder. Para calcular estas 'distancias químicas' recurrimos al coeficiente de Tanimoto (Haranczyk y Holliday, 2008) que permite comparar dos compuestos químicos generando un número que representa la similitud entre ellas. Un coeficiente de 1 representa dos moléculas que son idénticas, o cuyas diferencias

estructurales son tan sutiles que el método que utilizamos para calcular las distancias no es capaz de detectar (ver detalles en Métodos). Pero para poder interpretar los resultados debemos saber primer como se distribuyen las similitudes entre todos los compuestos de nuestra muestra analizada para saber si los compuestos son más o menos diferentes dentro de esta 'población química' (figura 42).

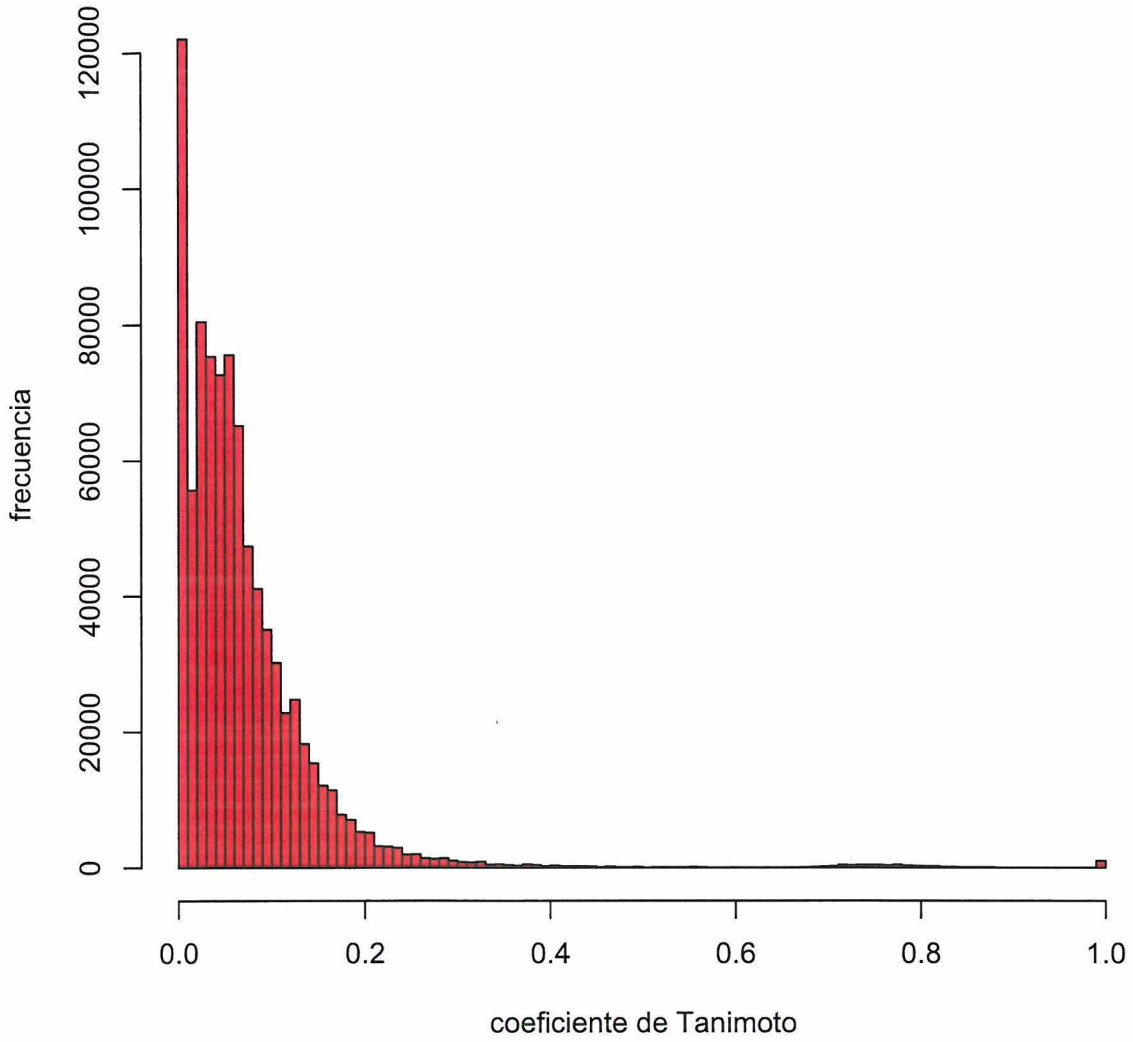



**Figura 42.** Distribución de los valores de similitud entre compuestos químicos implicados en procesos de biodegradación almacenados en la base de datos Bionemo. Los valores de similitud están calculados utilizando el coeficiente de Tanimoto.

La media de la distribución es 0,075 y la mediana 0,06. La mediana puede orientarnos mejor ya que la distribución no es una distribución normal, lo que comprobamos realizando el test de Kolmogorov-Smirnov para el que obtuvimos un valor p menor de 0,001. Hay que destacar que aproximadamente sobre el 0,3 de similitud nos alejamos de la mayoría de los valores de similitud de la población, así que consideramos una similitud

mayor de 0,3 es muy alta para esta población química. Una similitud de 0,06 representa la mediana de la población por lo que pones un umbral para valores de similitud por debajo de 0,06 en el que consideraremos que los compuestos son diferentes entre sí.

Teniendo en cuenta estos detalles podemos analizar las distancias entre los diferentes compuestos que actúan como efectores de un mismo regulador. En la figura 43 están representadas las mencionadas distancias.

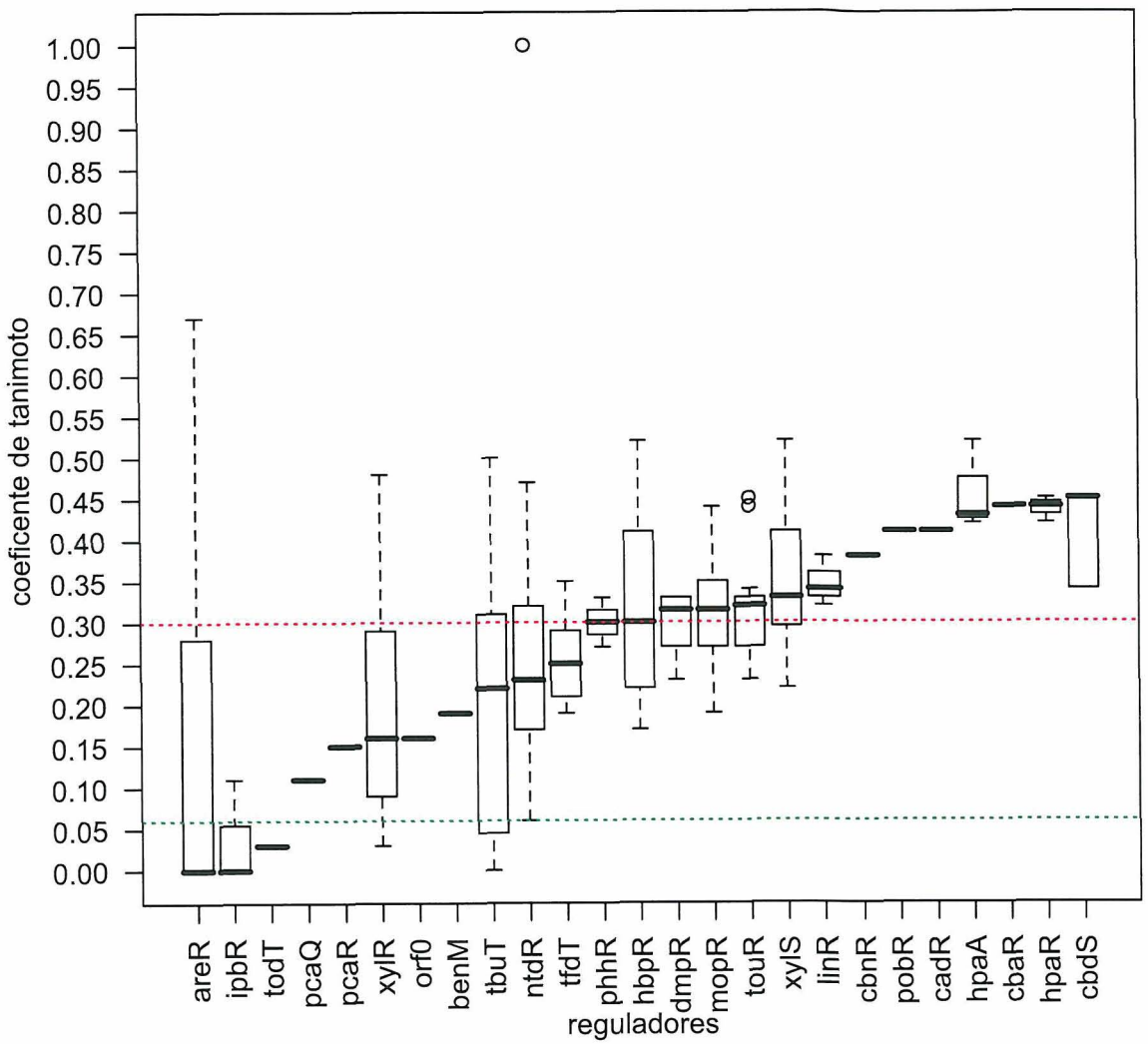



**Figura 43.** Distancia entre los efectores de un mismo regulador. Cada caja representa la distribución de las distancias entre moléculas efectoras reconocidas por un regulador. Están representados todos los reguladores de biodegradación que sabemos que responden ante más de un efector lo que nos permite calcular la distancia entre ellos. La línea roja representa la distancia 0,3 que podemos considerar como el límite a partir del cual dos compuestos tienen similitud alta. La línea verde representa la distancia 0,06 que es la mediana de la distribución de todos los compuestos que tenemos guardados en nuestra base de datos relacionados con biodegradación

Podemos destacar que 17 reguladores de un total de 25 analizados, el 68%, responden a efectores que tiene un coeficiente de Tanimoto menor de 0,3, es decir, son capaces de responder ante efectores diferentes entre sí, siempre hablando de diferencias dentro de esta población química. Hay 5 reguladores, un 20%, que presentan distancias menores de 0,06. Esto significaría que el 20% de los reguladores son capaces de responder a efectores que podemos considerar bastante diferentes entre sí ya que no alcanzan ni siquiera la similitud de la mediana de la población química, siendo la población de por sí bastante heterogénea. Entre estos reguladores encontramos, por ejemplo, a IpbR de *Pseudomonas putida* RE204 que es capaz de interaccionar con compuestos tan diversos como tricloroetileno, naftaleno o p-cimeno (Selifonova y Eaton, 1996). Por otro lado, 8 reguladores, un 32%, son muy específicos y sólo responden a compuestos con un coeficiente de Tanimoto mayor de 0,3, es decir, muy parecidos entre sí. En esta categoría estaría, por ejemplo, CadR de *Bradyrhizobium sp.* HW13 que imteacciona con 2,4-dicloro-fenoxiacetato y 4-cloro-fenoxiacetato (Kitagawa *et al.*, 2002). Podemos mencionar otro ejemplo: CbaR de *Conidiobolus coronatus* BR60, que interacciona con 3,4-dihidroxi-benzoato y 3-hidroxi-benzoato (Providenti y Wyndham, 2001). Esta variedad de respuestas de los diferentes reguladores se ajusta a lo predicho por la teoría del 'ruido regulatorio' ya que existen reguladores muy flexibles que serían los que están integrando nuevas señales y otros más específicos que podrían haber refinado su respuesta según sus necesidades.

**Un mismo efector puede interactuar con distintas familias de reguladores**

Si hay reguladores que son específicos y otros que son promiscuos, ¿de qué depende esa especificidad? Para intentar responder a esa pregunta buscamos patrones que nos indiquen si existe alguna relación conservada entre proteínas y compuestos químicos. La primera aproximación en esta dirección la hacemos comparando las familias de reguladores con los efectores (figura 44). Intentamos saber si proteínas parecidas entre sí, por ejemplo que pertenezcan a una misma familia de reguladores, se inducen por compuestos similares.

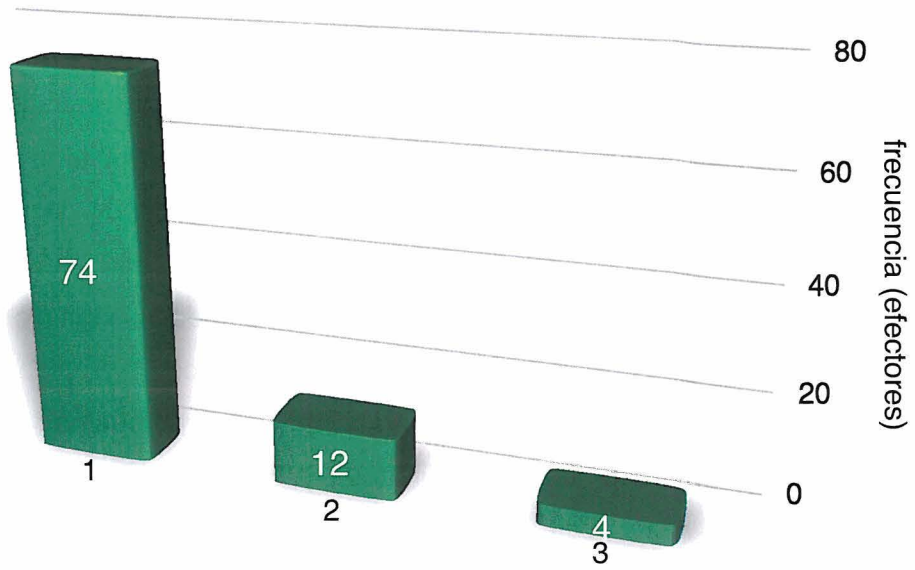



**Figura 44.** Número de familias que interactúan con un mismo efector. En la figura se representa en el eje de abscisas el número de familias diferentes con las que interactúa un mismo compuesto y en las ordenadas la frecuencia con la que esto ocurre

En la gráfica se observa que lo más habitual es que un compuesto interactúe con una única familia de reguladores. Esto era de esperar dado que la mayoría de compuestos solo son efectores de un único regulador. Podemos afinar un poco más este análisis calculando el número de reguladores que interactúan con un mismo efector y a cuantas familias pertenecen. El resultado se muestra en la figura 45. Aquí se observa que no parece haber una conservación entre los efectores y las familias de reguladores ya que en los casos de efectores que interaccionan con varios reguladores estos reguladores con frecuencia pertenecen a distintas familias. Es destacable el caso del fenol que interactúa con 8 reguladores distintos todos pertenecientes a la familia XylR de reguladores transcripcionales ya que en este caso si que parece haber una conservación entre las proteínas reguladoras y el compuesto inductor. Es diferente el caso del benzoato que es el efector de 5 reguladores que pertenecen a 3 familias distintas. Entre ellos están BenM de *Acinetobacter sp.* ADP1 que pertenece a la familia LysR (Clark *et al.*, 2004), PcaR de *Pseudomonas putida* PRS1, que pertenece a la familia IclR (Romero-Steiner *et al.*, 1994), y XylS del plásmido TOL en *Pseudomonas putida* mt2, que pertenece a la familia AraC (Ramos *et al.*, 1986). Otro ejemplo destacable es el del tolueno que interacciona con 6

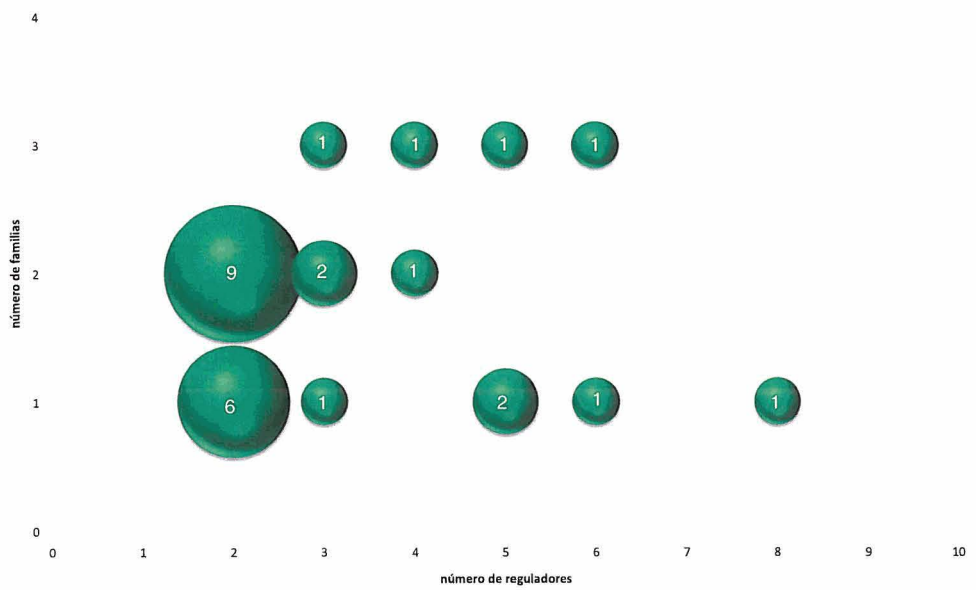



**Figura 45.** Número de reguladores y familias de reguladores con los que interactúa un efector en biodegradación. Hay que tener en cuenta que no hemos representado los efectores que interactúan con un único regulador, que son 63, ya que no resultan informativos para este análisis y hacen que el resto del gráfico sea más difícil de ver. El eje de abscisas representa el número de reguladores con los que interactúa un efector y el de ordenadas el número de familias a las que pertenecen los mencionados reguladores. El tamaño de las bolas es el número de efectores en esa categoría.

reguladores distintos que pertenecen a 3 familias o el 3,4-dihidroxi-benzoato que interactúa con tres reguladores de tres familias distintas.

Concluyendo, no parece que un efector interactúe con más frecuencia con la misma familia de reguladores. Esto podría sugerir la idea de que no tiene que existir necesariamente una correlación entre similitud de compuesto y similitud de regulador.

## Reguladores similares no tienen porqué ser inducidos por compuestos similares

Para contrastar la afirmación de que no tiene que existir necesariamente una correlación entre similitud de compuesto y similitud de regulador comparamos el parecido entre secuencias de reguladores con el que existe entre los compuestos que los inducen. Esta es la mejor forma de comprobar si realmente reguladores similares son inducidos por compuestos similares o si la inducción por un tipo de compuesto es independiente del tipo de regulador. Para calcular la similitud entre compuestos utilizamos de nuevo el coeficiente de Tanimoto. La gran cantidad de datos a analizar si representamos todas las distancias químicas frente a todas las identidades entre reguladores hace que sea más razonable dividir los datos en categorías para poder interpretarlos. Formamos tres grupos,

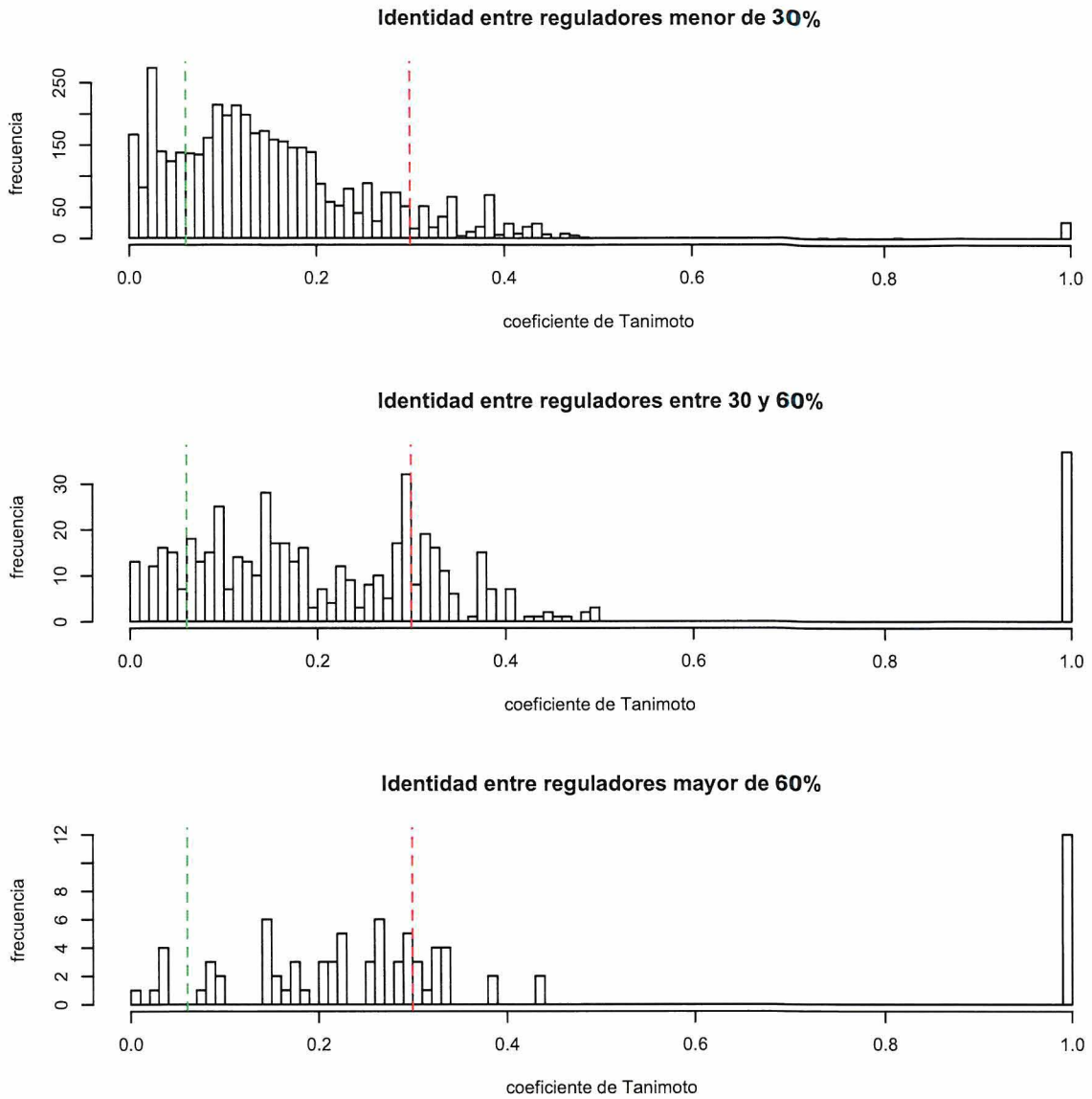



**Figura 46.** Distribuciones de las distancias químicas entre los inductores de distintos reguladores. Los comparaciones están clasificadas en 3 grupos según la identidad de secuencia entre los reguladores. El gráfico de arriba representa la distribución de coeficientes de Tanimoto calculados entre inductores de reguladores que comparten menos de un 30% de identidad de secuencia. El gráfico de en medio representa las comparaciones entre reguladores con una identidad de secuencia entre 30% y 60%. El gráfico de abajo representa las comparaciones entre reguladores con más de un 60% de identidad de secuencia. En los tres gráficos la línea verde de puntos representa la mediana de las distancias entre compuestos de toda la población química almacenada en Bionemo. La línea roja de puntos representa el coeficiente de Tanimoto de 0,3 a partir del cual consideramos los compuestos muy similares dentro de esta población química. Conviene señalar que el eje de ordenadas cambia en cada gráfico, disminuyendo el rango representado, mayor en el gráfico representado más arriba que en el que tiene debajo, ya que existen muchos más comparaciones entre reguladores con poca identidad de secuencia entre ellos que entre reguladores más parecidos entre sí.

el primero con las distancias químicas entre inductores que comparten menos de un 30% de identidad de secuencia. El segundo grupo incluye las distancias entre reguladores que comparten más de un 30% pero menos de un 60% de identidad entre sus secuencias. El

tercer grupo es el de las distancias entre reguladores que comparten más de un 60% de identidad (figura 46). Según un estudio sobre identidad de secuencia y transferencia de función en proteínas se puede asumir que si existe más de 60% de identidad entre sus secuencias se mantiene la especificidad de sustrato entre dos proteínas. A más de 40% de identidad entre secuencias se mantiene la actividad catalítica. Finalmente con un 30% de identidad se mantiene la estructura y se considera que tienen un origen evolutivo común (Devos y Valencia, 2000). De este modo, interpretamos que los reguladores de la categoría de menos del 30% de identidad de secuencia no tienen un origen evolutivo común. Los que tienen menos de 60% de identidad si lo tienen pero no deberían compartir especificidad de sustrato. Finalmente, los que comparten más de un 60% deberían tener especificidad de sustrato similar. En las gráficas de la figura 46 se observa como la similitud entre compuestos inductores va aumentando según aumenta la identidad de secuencia como era de esperar. A pesar de ello, encontramos que para los reguladores con poca identidad de secuencia tenemos un número considerable de compuestos por encima del límite de 0,3 de similitud, lo que interpretamos como compuestos muy parecidos entre sí dentro de esta población química. De hecho existe una cantidad reseñable de compuestos idénticos (coeficiente de Tanimoto 1). Observemos el otro extremo, la comparación entre inductores de reguladores con más de un 60% de identidad. Aunque encontramos una proporción mucho mayor de compuestos similares también encontramos que reguladores muy similares son inducidos por compuestos muy distintos entre sí: por debajo de la mediana de toda la población química e incluso con coeficiente de Tanimoto de 0.

En resumen, todo parece indicar que compuestos similares no tienen que inducir necesariamente reguladores similares y viceversa. Una explicación a esto podría ser que la inespecificidad inicial de los reguladores que, según la teoría del ruido regulatorio, les permite responder a nuevas señales no tiene su origen en ninguna familia de reguladores en concreto. De este modo, la respuesta a nuevos compuestos aparece de forma independiente en distintos reguladores de distintas bacterias que acaban realizando la misma función con distintos elementos en lo que seria un caso de convergencia evolutiva.

### En los operones catabólicos de biodegradación hay mucha más inducción 'gratuita' que en *Escherichia coli*

El reconocimiento del inductor por el regulador es el primer paso en la integración de la respuesta a nuevas señales. Pero para que la respuesta a la señal sea útil, esta debe

estar relacionada con los procesos de degradación que va a desencadenar la presencia del efector. En otras palabras, lo más deseable es que la presencia de una molécula provoque la expresión de los genes que van a degradar esta molécula. Para comprobar el nivel de coordinación entre la inducción de la expresión de los promotores y la especificidad de las enzimas expresadas comparamos la molécula efectora con los sustratos degradados por estas enzimas. Si el efector es un sustrato de alguna reacción realizada por las enzimas lo clasificamos como coordinado, si no lo clasificamos como gratuito. Llamamos inducción gratuita a la expresión de un operón que codifica para enzimas que no degradan la molécula efectora. Comparamos las proporciones de expresión coordinada y gratuita en biodegradación frente a las de *Escherichia coli* (figura 47).

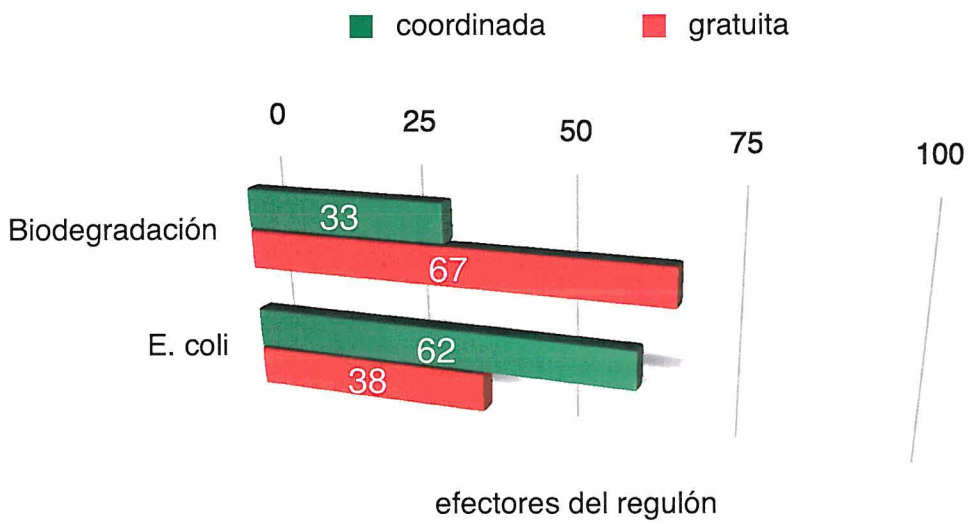



**Figura 47.** Clasificación en porcentajes de los efectores de los regulones en biodegradación y en *Escherichia coli*. Si el efector del regulador es sustrato de alguna de las enzimas expresadas por los promotores activados por la acción del regulador clasificamos la inducción como coordinada, si no la clasificamos como gratuita

En la figura se observa que en biodegradación existe una proporción mucho mayor de inducción gratuita que de coordinada lo que indica una regulación mucho menos específica. Existen diferencias significativas entre las proporciones de los dos tipos de inducción entre los sistemas de biodegradación y *Escherichia coli* (ji-cuadrado, $p < 0{,}001$). Podemos afirmar que la proporción de inducción gratuita en biodegradación es mayor. También es interesante mencionar que en el caso de los sistemas de biodegradación en la mayoría de la inducción coordinada el inductor es el sustrato de la primera enzima

codificada en el operón. Esto se ha sugerido que puede servir para hacer que los sistemas tengan mayor robustez y capacidad de respuesta (Wall *et al.*, 2004).

**Las operones pueden ser inducidos por compuestos muy diferentes a los sustratos de las enzimas codificadas en ellos**

Sabemos que existe una gran proporción de inducción gratuita en biodegradación pero, ¿hasta que punto es gratuita?¿se puede cuantificar este nivel de 'gratuidad'? Sí, podemos comparar el efector con los sustratos de las enzimas codificadas por el regulón y, de esta manera ver hasta que punto es diferente el efector a los sustratos de las enzimas. El resultado se muestra en la figura 48.

En la figura se puede observar que la mayoría de los inductores tiene un gran parecido con alguno de los sustratos de los degradados por los complejos enzimáticos expresados en respuesta a su presencia, que estaría representado por los valores por encima de la línea roja de puntos de la figura. Por otro lado, también encontramos sustratos muy diferentes de los inductores, lo que estaría representado especialmente por los valores

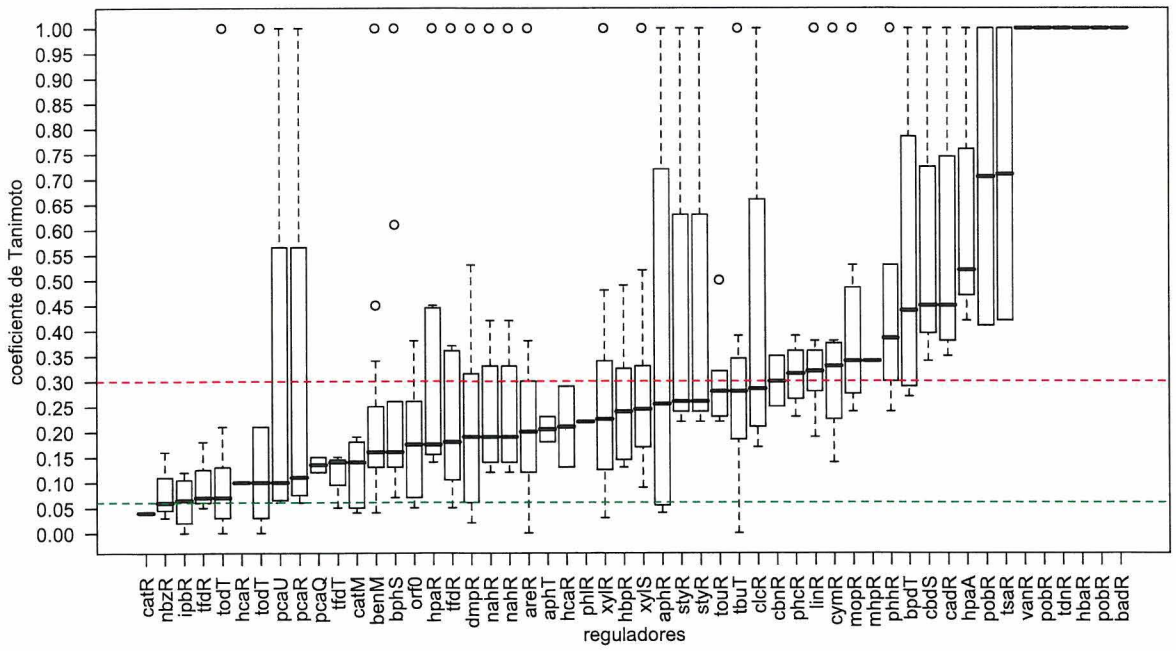



**Figura 48.** Distribuciones de las distancias entre los efectores de un regulador y los sustratos degradados por las enzimas expresadas por su regulón. Cada caja representa las distancias entre los compuestos asociados a un regulador. La línea roja de puntos representa la distancia entre compuestos de 0,3 a partir de la cual consideramos que los compuestos son muy parecidos entre sí para esta 'población química' . La línea verde de puntos representa la mediana de la población de distancias químicas

por debajo de la línea verde. Las distribuciones abarcadas por cada caja son muy amplias pero esto es razonable, ya que estamos comparando el inductor con todos los sustratos y, aunque todos sean similares entre sí, van sufriendo transformaciones por las reacciones y algunos pueden acabar siendo muy distintos. Para afinar el análisis filtramos las distancias entre los inductores y nos quedamos únicamente con la que se corresponde con el sustrato más parecido al inductor. El resultado se muestra en la figura 49.

Ahora las distribuciones son menos amplias pero incluso eligiendo el sustrato más parecido al efector existen reguladores en los que la mayoría de los efectores son muy distintos de los sustratos de los complejos enzimáticos expresados. Parece que los sistemas regulatorios de biodegradación son muy tolerantes al ruido y puede que esta flexibilidad les haya servido para integrar nuevas señales. Un ejemplo sería el caso de

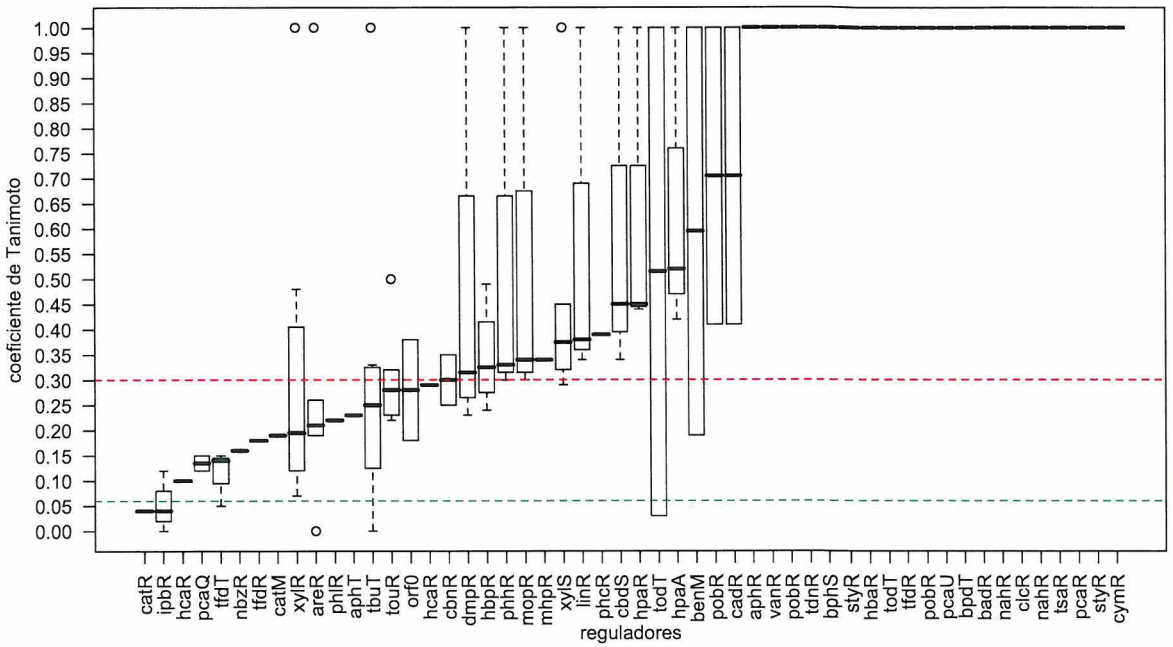



**Figura 49.** Distribuciones de las distancias entre los efectores de un regulador y el sustratos degradado por las enzimas expresadas por su regulón más parecido al efector con el que se compara. Cada caja representa las distancias entre los compuestos asociados a un regulador. La línea roja de puntos representa la distancia entre compuestos de 0,3 a partir de la cual consideramos que los compuestos son muy parecidos entre sí para esta 'población química' . La línea verde de puntos representa la mediana de la población de distancias químicas

CatR de *Pseudomonas putida*. El inductor de CatR es el cis-cis-muconato y es químicamente muy diferente del fenol y el catecol que son los sustratos de las enzimas que se expresan en su presencia. Curiosamente, en este caso el producto de una de la enzima que degrada catecol es cis,cis-muconato, que no hemos comparado ya que sólo comparamos los sustratos. El cis,cis-muconato es muy diferente de los otros compuestos

porque no contiene un anillo aromático. Otro ejemplo sería IpbR de *Pseudomonas putida* RE204 del que ya hablamos anteriormente al comentar que tiene un amplio rango de respuesta a distintos inductores. Era lógico esperar que si es inducido por compuestos diferentes entre sí estos compuestos puedan ser también diferentes de los sustratos de la ruta catabólica que inducen. Como conclusión, podemos afirmar que los operones catabólicos pueden ser expresados en respuesta a compuestos diferentes de los sustratos de las enzimas codificadas en ellos en un fenómeno que hemos llamado inducción gratuita. Esta inducción gratuita llega hasta el punto de que en la mayoría de los sistemas que hemos estudiado el inductor no es el sustrato de ninguna de las enzimas. Por otra parte, también en la mayoría de los sistemas hay al menos un inductor muy parecido a alguno de los sustratos de la ruta, preferentemente el primer sustrato, lo que hace que los circuitos de regulación sean más robustos y tengan una mayor capacidad de respuesta.

**Las enzimas y los reguladores evolucionan de forma independiente**

Operones catabólicos diferentes, que degradan diferentes compuestos, pueden usar mecanismos reguladores similares. Un ejemplo sería el caso del plásmido TOL, que degrada tolueno , y el operón *dmp* del plásmido pVI150, que degrada fenol, ambos con un sistema regulatorio similar (Cases y de Lorenzo, 2005). También ocurre que operones catabólicos similares pueden estar regulados por reguladores distintos en distintas especies como es el caso de la regulación del catabolismo del bifenilo. En *Rhodococcus sp.* M5, el operón catabólico *bpdC1C2BADE* está controlado por un sistema de dos componentes, BpdST, mientras que en *Pseudomonas sp.* KKS102 el regulador BphS pertenece a la familia GntR de represores. Se ha sugerido que los reguladores y las enzimas cuya expresión están controlando pueden evolucionar de forma independiente llamando a este fenómeno gato blanco/gato negro en referencia a que distintos mecanismos producen el mismo resultado (Cases y de Lorenzo, 2003). Para abordar esta cuestión hemos comprobado el parecido entre los reguladores que controlan operones homólogos. Para considerar dos operones como homólogos debían de cumplir unos criterios de identidad de secuencia entre sus genes: al menos dos tercios de sus genes debían de tener un cierto porcentaje de secuencia idéntico (ver detalles en Métodos). Seleccionamos tres porcentajes: 30, 60 y 90. Como hemos mencionado anteriormente, según un estudio sobre identidad de secuencia y transferencia de función en proteínas se puede asumir que si existe más de 60% de identidad entre sus secuencias se mantiene la

especificidad de sustrato entre dos proteínas. A más de 40% de identidad entre secuencias se mantiene la actividad catalítica. Finalmente con un 30% de identidad se mantiene la estructura y se considera que tienen un origen evolutivo común (Devos y Valencia, 2000). En la tabla 2 se muestran los resultados obtenidos usando los distintos umbrales de identidad de secuencia para considerar los operones homólogos

| Porcentaje de identidad entre operones | Porcentaje de identidad entre reguladores | | | |
|---|---|---|---|---|
| | <30% | <60% | <90% | >90% |
| >30% | 72 | 23 | 4 | 8 |
| >60% | 64 | 6 | 4 | 8 |
| >90% | 64 | 4 | 2 | 7 |

**Tabla 2.** La tabla representa la similitud entre operones frente a la similitud entre los reguladores que los controlan. Por ejemplo, en 72 comparaciones entre operones que tenían más de un 30% de identidad de secuencia en al menos dos tercios de sus genes los reguladores no llegaban al 30% de identidad de secuencia entre ellos.

Se observa que, aunque comparemos operones muy parecidos entre sí, los reguladores no suelen ser similares. Esto apoyaría la idea de que la evolución de los operones catabólicos y de los reguladores ha ido por caminos separados.

### Los genes que son contiguos a sus operones regulados tienen cierta tendencia a estar más conservados

Anteriormente hemos observado que la mayoría de los operones catabólicos implicados en procesos de biodegradación presentan el gene del regulador contiguo a sus genes regulados. Se ha sugerido que la transferencia horizontal de genes podría estar implicada en la conservación de los genes de los reguladores que se encuentran adyacentes a sus genes regulados (Hershberg *et al.*,2005) y que estos reguladores 'contiguos' conservan su posición con respecto al operón regulado en otros genomas (Korbel *et al.*,2004). Para comprobar si esto podría ser así en los sistemas de biodegradación comparamos la similitud entre reguladores de operones con su regulador contiguo con la de operones con el regulador separado. Los resultados se muestran en las tablas 3 y 4.

| Porcentaje de identidad entre operones | Porcentaje de identidad entre reguladores contiguos | | | |
|---|---|---|---|---|
| | <25% | 25-50% | 50-75% | >75% |
| >30% | 1 | 4 | 3 | 3 |
| >60% | 1 | 0 | 0 | 3 |
| >90% | 1 | 0 | 0 | 1 |

**Tabla 3.** La tabla representa la similitud entre operones frente a la similitud entre los reguladores que los controlan para los operones que tienen reguladores contiguos. Por ejemplo, en 4 comparaciones entre operones que tenían más de un 30% de identidad de secuencia en al menos dos tercios de sus genes los reguladores tenían entre un 25 y un 50% de identidad de secuencia entre ellos.

| Porcentaje de identidad entre operones | Porcentaje de identidad entre reguladores separados | | | |
|---|---|---|---|---|
| | <25% | 25-50% | 50-75% | >75% |
| >30% | 11 | 6 | 0 | 2 |
| >60% | 11 | 4 | 0 | 2 |
| >90% | 11 | 3 | 0 | 2 |

**Tabla 4.** La tabla representa la similitud entre operones frente a la similitud entre los reguladores que los controlan para los operones que tienen reguladores separados. Por ejemplo, en 6 comparaciones entre operones que tenían más de un 30% de identidad de secuencia en al menos dos tercios de sus genes los reguladores tenían entre un 25 y un 50% de identidad de secuencia entre ellos.

Comparando los valores de las dos tablas se observa que en los operones que tienen reguladores contiguos encontramos una mayor similitud entre reguladores con valores más altos de identidad (tabla 3). En el caso de los operones con los reguladores separados la mayoría de los operones homólogos, independientemente del nivel de identidad de secuencia entre sus genes, tienen valores bajos de similitud entre reguladores. Estos datos parecen apoyar la idea de que la regulación se conserva más cuando los genes del regulador se encuentran adyacentes a los genes regulados y que se podrían transmitir junto con la regulación entre especies. De todas formas, debemos de ser cautos a la hora de sacar conclusiones ya que la escasez de datos nos obliga a ello.

Resumiendo, hemos visto que los reguladores implicados en biodegradación puede ser muy flexibles respondiendo a compuestos diferentes. Esto está asociado a la 'inducción gratuita' de los operones catabólicos en respuesta a compuestos efectores que nos son

sustratos de la ruta expresada, aunque generalmente son parecidos. También hemos observado que operones homólogos no tienen porqué ser regulados por reguladores homólogos, lo que sugiere una evolución independiente del metabolismo y la regulación. Finalmente, hemos detectado cierta tendencia a que la regulación esté más conservada cuando los genes regulados están adyacentes al gen del regulador.

# DISCUSIÓN

El vertido masivo de compuestos tóxicos por la actividad industrial humana es un hecho relativamente reciente. La aparición de bacterias capaces de actuar en respuesta a compuestos contaminantes es interesante desde varios puntos de vista. Por un lado, entender como funcionan los mecanismos que utilizan estas bacterias para degradar estos compuestos contaminantes nos puede servir para tratar de aplicar estos conocimientos para limpiar el medio ambiente. Por otro, estamos en situación de estudiar como se forman unos sistemas emergentes de regulación, lo que resulta muy interesante desde el punto de vista evolutivo. Durante el desarrollo de esta tesis, hemos intentado caracterizar y comparar de forma sistemática los sistemas de regulación de estas bacterias implicadas en biodegradación para intentar explicar sus propiedades en un contento evolutivo y funcional.

**Circuitos de regulación: activadores polivalentes, estabilidad y flexibilidad.**

Estudiando los tipos de reguladores que se encuentran en los sistemas de regulación implicados en biodegradación encontramos una mayoría de activadores. El origen de esta prevalencia de los activadores no puede ser explicado por un efecto fundador ya que pertenecen a distintas familias de reguladores. Independientemente de la familia a la que pertenecen, los reguladores son capaces de interactuar con un mayor número de efectores que sus equivalentes en *Escherichia coli.* Paralelamente a esta mayor cantidad de activadores encontramos una mayor proporción de promotores activables. Esto nos indica un caso de convergencia evolutiva ya que partiendo de diferentes orígenes se ha llegado a una misma solución. ¿Qué ventaja adaptativa puede ofrecer el regular por activadores para que esta haya sido la opción más seleccionada? Se ha sugerido que la selección del tipo de reguladores utilizados puede estar asociada a la demanda de la expresión de los genes (Savageau, 1977): en caso de que se necesite expresar unos genes con mucha frecuencia se seleccionará su expresión por activadores. Quizá el aumento de los compuestos tóxicos en el medio ambiente a sido un factor en esta selección a favor de los activadores. Por otro lado, en caso de mutación de que una mutación en el

regulador o en el sitio de unión impidiera su buen funcionamiento los genes catabólicos se expresarán sin control si están regulados por un represor. Esto no ocurriría si estuvieran bajo la influencia de un activador, lo que supondría otra posibilidad que favorecería la selección de activadores.

Otro factor a tener en cuenta es la polivalencia de los reguladores en función de la localización con respecto al inicio de transcripción de los sitios de unión al ADN. Ya se había descrito con anterioridad que un mismo regulador podía actuar como activador o represor si el sitio de unión al que se asocia está más lejos o más cerca del inicio de transcripción (Collado-Vides *et al.*, 1991). Así el mismo regulador puede activar la expresión de los operones catabólicos y reprimir la expresión de su propio gen uniéndose a dos sitios de unión que estén a una determinada distancia del inicio de transcripción. De hecho, así es como ocurre

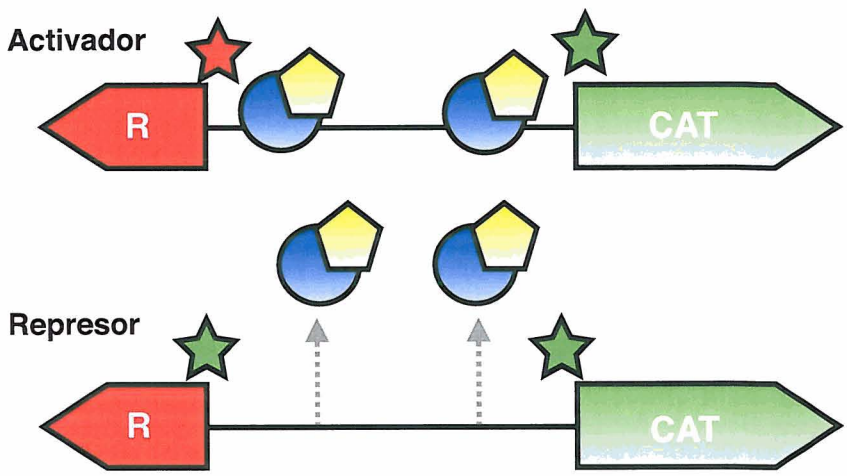



**Figura 50**. Control de los promotores de los genes catabólicos y del gen del regulador por medio de activadores, arriba, o represores, abajo. En respuesta a una molécula efectora (pentágono amarillo), el activador (círculo azul) se une al ADN a diferentes distancias del inicio de transcripción provocando la expresión (estrella verde) del operón catabólico (marcado como CAT) y la represión (estrella roja) de su propio gen (marcado como R) . En el caso del represor, en respuesta al efector el represor libera su unión del ADN causando la expresión tanto del promotor catabólico como la de su propio gen.

en la mayoría de los casos en los circuitos que hemos analizado: un mismo regulador activa el promotor de los genes catabólicos y reprime la expresión de su propio gen uniéndose a sitios de unión que están más o menos cerca del inicio de transcripción en repuesta a una molécula efectora (figura 50, arriba). Ejemplos de este tipo de circuito pueden ser XylR y el operón *xylUWCMABN* del plásmido TOL (Ramos *et al.*,1997) o CatR y el operón *catBC* de *Pseudomonas putida* (Parsek el al., 1992). Pero este tipo de control no es posible por medio de un

represor ya que, en respuesta al efector, libera su represión tanto del promotor catabólico como del gen del regulador (figura 50, abajo). Y así es como ocurre en biodegradación para los promotores catabólicos controlados por represores. Es el caso del represor HpaR de *Escherichia coli* que en presencia de 4-hidroxi-fenil-acético, 3-hidroxi-fenil-acético o 3,4-hidroxi-fenil-acético libera la represión tanto del promotor del operón catabólico que controla, *hpaGEDFHI*, como la que ejerce sobre el promotor de su propio gen (Galán *et al*, 2003). El represor PcaU de *Acinetobacter sp.* ADP1 que controla la expresión del operón *pcaIJFBDKCHG,* que degrada protocatechuato, responde a sus efectores de la misma manera que HpaR liberando la represión del promotor que expresa su propio gen junta a la del promotor catabólico (Gerischer *et al*., 1998). En este último caso se ha descrito un control adicional por regulación cruzada con los reguladores que controlan la degradación del catecol (Brzostowicz *et al*., 2003). Podría ocurrir que un represor cambiara su conformación y, en el caso de que el operón catabólico y el gen del regulador compartiera sus zonas de regulación, se desplazara por el ADN reprimiendo el gen del regulador y permitiendo la expresión del promotor catabólico. Esto requeriría un mecanismo de regulación más complejo y una organización genética determinada que no se requieren si el regulador es un activador. Pero, ¿qué ventaja puede ofrecer ser capaz de reprimir el propio gen del regulador a la vez que se activa el promotor catabólico? Según estudios teóricos sobre la estabilidad, robustez y capacidad de respuesta de los circuitos de regulación, para que un sistema inducible sea más estable el gen del regulador debe reprimirse al inducirse el promotor catabólico (Wall *et al*., 2004). También se asegura en estos estudios que este tipo de autorregulación permite una mayor capacidad de respuesta y hace que el sistema sea más robusto. De esta forma, los sistemas de regulación controlados por activadores son más estables sin necesidad de un control adicional de la regulación y la selección actuaría al nivel del circuito más que al nivel de los componentes, como se ha sugerido anteriormente (Cases y de Lorenzo, 2005). En *Escherichia coli* el sistema más frecuente es el de la represión de los operones catabólicos. Quizá estos sistemas al estar más integrados en la maquinaria de regulación global del organismo no tienen tantas presiones selectivas que les obliguen a ser estables de forma independiente. A esto se añade que el tener los genes controlados por un menor número de reguladores y expresados desde un menor número de promotores podría también redundar en una mayor estabilidad de los sistemas al tener por un

lado menos requerimientos para su inducción o inhibición y por otro menos promotores que controlar.

Apoyando esta relación lógica entre los componentes que forman los circuitos de regulación encontramos una relación física. Hemos observado que los genes de los reguladores que están controlando estos operones catabólicos se encuentran frecuentemente contiguos y transcritos de forma divergente al operón catabólico que controlan. En *Escherichia coli* también se había observado una situación similar (Warren y Wolde, 2004) en la que se describía que los operones que regulan unos a otros y los co-regulados tienden a presentar esta misma organización genética. Esta disposición permite que las zonas de regulación de los operones se solapen lo que permite añadir un nivel de control regulatorio adicional coordinando la expresión de lo operones. En este artículo se sugería que el control regulatorio puede ejercer una presión selectiva que favorece esta organización. Por otra parte, se ha descrito que los genes que responden a estímulos externos también se encuentran próximos en el genoma lo que les ayuda a comportarse como módulos de regulación que proporcionan una respuesta rápida a variaciones medioambientales de una forma coordinada (Janga *et al.*, 2007). Todas estas propiedades puestas en común suponen una ventaja para los circuitos de regulación de las bacterias implicadas en biodegradación.

**Integración con la fisiología y selección a nivel de circuitos**

En los sistemas de biodegradación encontramos una mayor presencia de promotores asociados a sigma 54 que en *Escherichia coli*. Esta mayor presencia no está causada por un efecto fundador ya que los promotores asociados a sigma 54 expresan grupos de genes no homólogos. El papel ejercido en la integración con la fisiología del microorganismo que se ha descrito para los promotores asociados a sigma 54 (Carmona *et al.*, 1997; Macchi *et al.*, 2003; Van Dien y de Lorenzo, 2003) podría ser la causa de esta diferencia. Los reguladores implicados en biodegradación son capaces de responder a compuestos muy diferentes entre sí y diferentes de los sustratos de la ruta que controlan, algo que discutiremos en detalle más adelante. También hemos visto que para expresar un promotor catabólico no se requiere un tipo concreto de regulador ni de sitio de unión. El disponer de un sistema tan flexible puede tener un efecto negativo si no está bien

integrado con la fisiología del organismo, ya que si no se los genes catabólicos no se expresan cuando se necesitan o si se expresan de forma descontrolada puede suponer una desventaja adaptativa por el derroche de energía que causarían. También se ha descrito que la integración de los promotores catabólicos de biodegradación por diferentes mecanismos puede parecer una estrategia poco adecuada desde el punto de vista de la ingeniería del sistema ya que tener muchos mecanismos distintos para hacer el mismo trabajo podría parecer un derroche de recursos (Cases y de Lorenzo, 2005). Pero, como allí mismo se sugiere, lo que en una primera impresión podría parecer una inconveniente puede convertirse en una ventaja si se observa desde el punto de vista de una comunidad de microorganismos. Esta flexibilidad permite que estos operones sean integrados y utilizados por organismos en circuitos regulatorios muy diferentes utilizando los elementos disponibles para cada microorganismo. Los genes catabólicos que degradan xenobióticos se encuentran frecuentemente asociados a elementos móviles (Top y Springael, 2003). De esta forma, estos genes catabólicos se pueden extender lo que acabará resultando beneficioso para toda la comunidad microbiana. Se podría argumentar que un único organismo podría beneficiarse de esta capacidad y que no existe una presión selectiva sobre la comunidad. Quizá la supervivencia de una mayor variedad de microorganismos puede ofrecer una ventaja adaptativa a la comunidad por hacerla más resistente a diferentes amenazas que podrían acabar más fácilmente con una población formada por una única especie. En definitiva, la flexibilidad de los mecanismos de regulación sugiere que su selección actúa más que al nivel de los elementos que los componen, al nivel del circuito de regulación, como ya ha sido sugerido anteriormente (Cases y de Lorenzo,2005), y sugerimos que podría incluso actuar al nivel de la comunidad microbiana.

**Organización genética**

Como ya hemos comentado anteriormente, los genes implicados en procesos catabólicos de biodegradación se encuentran con frecuencia asociados a elementos móviles como plásmidos y transposones (Top y Springael, 2003). También se han descrito casos en los que conjuntos de genes que realizan procesos catabólicos similares aparecen en bacterias que no están directamente relacionadas filogenéticamente lo que se explica como eventos de transferencia

horizontal de genes (Furukawa y Fujihara, 2008). Esto parece indicar que estos genes pueden ser adquiridos por diferentes microorganismos por los procesos habituales de adquisición de fragmentos de ADN utilizados por las bacterias como la conjugación o la transformación (Thomas y Nielsen, 2005; Sørensen *et al.*, 2005). Si esto es cierto, la organización genética tanto de los genes implicados en el catabolismo como en los genes de los reguladores podría haber influido en el éxito de estas transferencias, entendiendo como éxito la transferencia de capacidades funcionales (complejos enzimáticos y rutas catabólicas). Por nuestra parte, hemos observado que los operones que contienen genes catabólicos implicados en biodegradación contienen un número más grande de genes que sus equivalentes en *Escherichia coli*, una media de seis genes por operón en biodegradación frente a los escasos 3 genes por operón de *Escherichia coli*. Comparando la forma en que se codifican los complejos en los operones hemos observado que excepto en muy raros casos los genes que codifican un complejo enzimático están asignados a un mismo operón. En estos raros casos, los genes se agrupan unos junto a otros en lo que se a llamado "uber-operones" (Lathe *et al.*, 2000) o "vecindarios de genes" (Rogozin *et al.*, 2002) , ya que se encuentran anexos aunque no se expresen desde el mismo promotor. Más aún, los genes que codifican un mismo complejo suelen estar agrupados unos junto a otros y los complejos que realizan reacciones consecutivas también suelen encontrarse codificados de forma consecutiva en los operones siguiendo el sentido de la transcripción. Todo esto ocurre con más frecuencia de lo que sería esperable por azar lo que indica que debe haber sido seleccionado por alguna razón. Existen tres razones que pueden explicar la conservación de los operones en las bacterias: primera, divergencia reciente; segunda, transferencia horizontal de genes reciente; y tercera, fuertes restricciones regulatorias o estructurales que seleccionan en contra la reorganización de los operones y conjuntos de genes (Tamames, 2001). Se han propuesto varias explicaciones para este fenómeno y se han propuesto modelos que se dividen en tres categorías: (1) los que consideran los beneficios de transcribir y traducir los genes agrupados en un espacio restringido dentro de la célula; (2) los que hablan de la amplificación de los genes como mecanismo para adquirir la co-regulación de genes; y (3) los que tienen en cuenta el efecto de la proximidad de los genes en la frecuencia de la transferencia coordinada de los genes y la recombinación. Dentro de este último modelo se han propuesto dos modelos que están relacionado con nuestro caso

de estudio. El modelo de 'coadaptación de Fisher' está basado en el concepto de coevolución: los genes cuyos productos están implicados en la misma ruta metabólica, especialmente si forman parte del mismo complejo enzimático, deberían evolucionar de forma coordinada (Townsend *et al.*, 2003). El otro modelo es el del ' operón egoísta', que sugiere que el agrupamiento en genes es beneficioso para los genes en el operón pero no necesariamente para el organismo que los alberga. La combinación de genes del operón sería seleccionada si confiere un fenotipo seleccionable (Lawrence, 1999). Lo que observamos en los operones que hemos estudiado está de acuerdo con ambos modelos. Los genes que codifican para las mismas rutas y los mismos complejos se agrupan en los mismos operones, como predice el modelo de coadaptación, . Por otro lado, la adquisición de complejos enzimáticos y complejos conectados confiere un fenotipo seleccionable, como predice el modelo del 'operón egoísta': al adquirirse estos genes, por alguno de los procesos de transferencia anteriormente mencionados, se han seleccionado en el organismo huésped los grupos de genes que le han permitido adquirir alguna nueva funcionalidad como puede ser el ser capaz de degradar algún compuesto, un complejo enzimático completo, o la adquisición de una ruta catabólica completa, varios complejos conectados. Esta última posibilidad sería la más seleccionada lo que explica la organización en operones largos que codifican varios complejos que realizan los sucesivos pasos de degradación de un compuesto. Quizá el adquirir la capacidad de degradar parcialmente un compuesto tóxico pueda estar seleccionada en el caso de que el producto generado reduzca su toxicidad. Además, si se adquiere la capacidad de degradar un compuesto tóxico hasta convertirlo en uno asimilable por el metabolismo central sí se obtiene una ventaja adaptativa que puede suponer una diferencia entre la supervivencia o no en situaciones en las que ese compuesto sea la única fuente de energía disponible. Esto podría explicar la organización en operones largos pero no indica que los genes tengan que estar necesariamente ordenados en el operón. La selección de este orden podría estar relacionada con el hecho de que favorece la adquisición de un mayor número de complejos completos y genes ordenados como comprobamos por medio de una simulación de los eventos de transferencia horizontal de genes. La presencia de los genes catabólicos en elementos móviles y el proceso de adquisición realizado por los distintos organismos que los contienen podría haber acelerado el proceso de selección de este orden (figura 51).

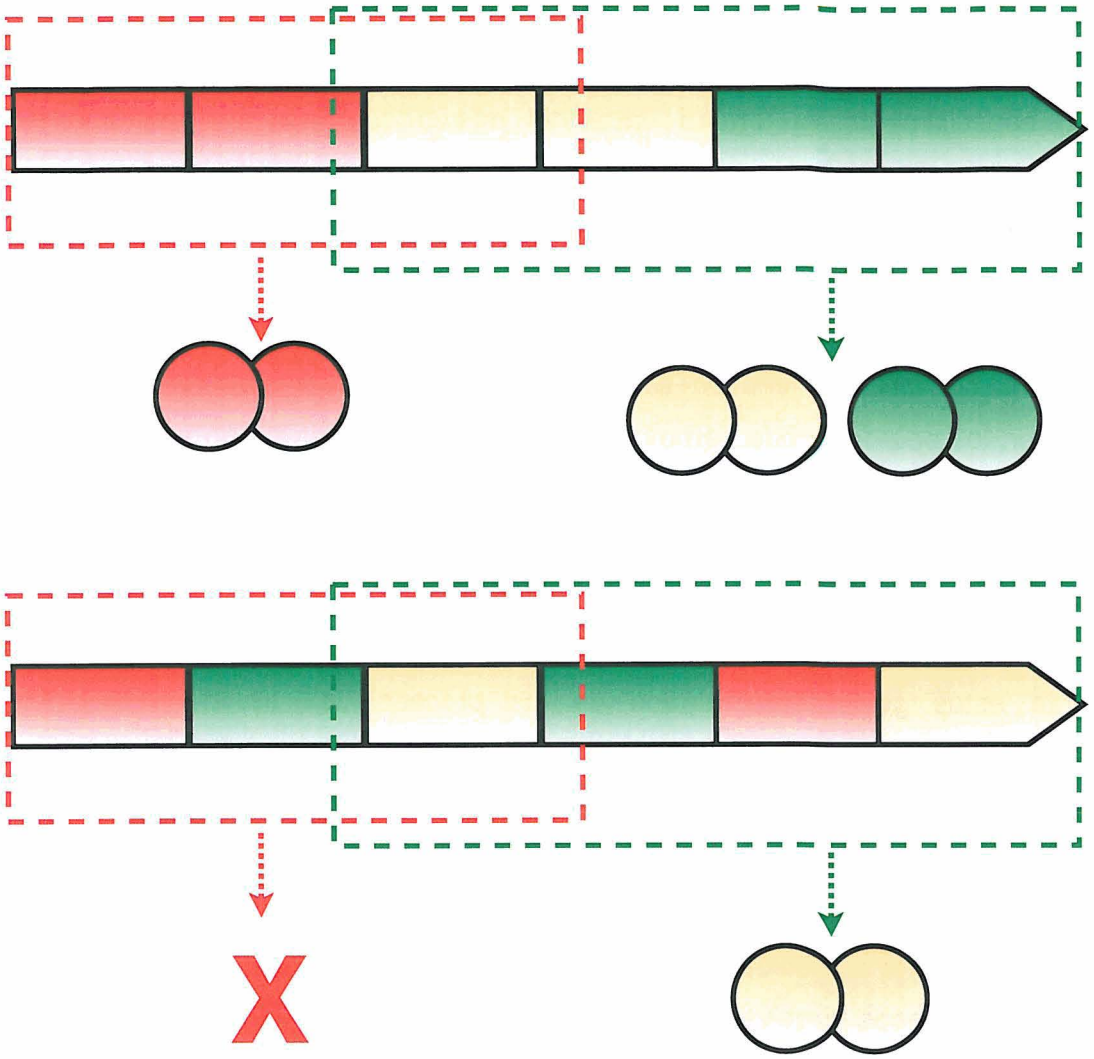



**Figura 51.** Simulación de transferencia comparando un operón con genes ordenados por complejos (arriba) frente a un operón con los genes desordenados. Los genes que codifican para el mismo complejo son del mismo color. Cada evento de transferencia está representado por una línea de puntos (roja o verde). Por ejemplo, en el operón ordenado (arriba) la transferencia representada por la línea de puntos verde permite la adquisición de dos complejos completos. Si los dos complejos estuvieran conectados se transmitiría una pequeña ruta metabólica. Si el compuesto degradado por el complejo naranja fuera tóxico se conseguiría lograr su detoxificación si el último paso de esta detoxificación lo realiza el complejo verde. La misma transferencia con el orden alternativo sólo permite adquirir un complejo que, si seguimos con el supuesto del compuesto tóxico, no elimina la amenaza tóxica.

A la existencia de esta organización genética se suma el agrupamiento de los genes en operones largos y ordenados que permite el control de la expresión de rutas catabólicas completas con una respuesta rápida, coordinada y constituyendo un módulo independiente todo ello ocupando el mínimo espacio posible en el ADN. De esta manera una bacteria que adquiriera este fragmento de ADN en un evento de transferencia horizontal de genes obtendría junto a la

capacidad catabólica la posibilidad de controlarla por medio de un circuito de regulación completo lo que podría suponer una clara ventaja adaptativa que podría ser seleccionada a favor. Esto también estaría de acuerdo con la sugerencia de que en los sistemas de biodegradación la selección actúa a nivel de circuitos de regulación más que a nivel de los componentes (Cases y de Lorenzo,2005) ya que en este caso estaríamos seleccionando el fragmento de ADN que incluye todos los componentes que van a formar el circuito regulador cuando se expresen los genes .

**Integración de nuevas señales**

Se ha sugerido que las bacterias implicadas en biodegradación adquieren la capacidad de respuesta a nuevas señales a través de un fenómeno llamado 'ruido regulatorio'. Según esta teoría los reguladores y promotores en un principio serían muy poco específicos permitiendo así responder a un gran número de compuestos distintos. Posteriormente, según fuera necesario, se irían haciendo más específicos suprimiendo las señales que no son deseables y ajustando la respuesta a los compuestos químicos de interés (de Lorenzo y Pérez Martín, 1996). El hecho de que los reguladores en los sistemas de biodegradación sean mucho menos específicos que los de los sistemas catabólicos de *Escherichia coli* parece apoyar la primera parte de esta teoría, la adquisición de nuevas señales a partir de la inespecificidad. Hemos comprobado que estos reguladores no sólo son capaces de responder a diferentes compuestos si no que estos compuestos pueden llegar a ser bastante diferentes entre sí. Esto tiene como consecuencia una mayor inducción gratuita de los operones en biodegradación que en *Escherichia coli*, considerando como inducción gratuita la expresión de operones en respuesta a una señal que no es un sustrato de la ruta catabólica expresada por el operón. Los compuestos inductores pueden ser muy diferentes de los sustratos de la ruta que inducen, tanto dentro de cada circuito regulatorio como entre circuitos, lo que podría estar relacionado con el ajuste de la respuesta a una señal más específica que se iría seleccionando cuando fuera necesario. Con respecto a este ajuste, cabe destacar que cuando el inductor coincide con el sustrato suele ser el primer paso de las reacciones expresadas por el operón, lo que según Savageau hace que la respuesta sea más rápida y robusta. Si el inductor coincide con un compuesto intermedio de la ruta catabólica se pierde

robustez y capacidad de respuesta pero se gana estabilidad (Wall *et al.*, 2004). En este tipo de circuitos parece que no es necesario reducir la capacidad de repuesta y robustez en aras a conseguir una mayor estabilidad porque esta se puede alcanzar a través de la autorepresión del gen del regulador. Así todo parece indicar que la poca especificidad de los reguladores permite captar en primera instancia la señal para ejercer una respuesta. También hemos encontrado reguladores muy específicos, que podrían ser el resultado de la posterior selección de una respuesta más rápida, estable y robusta.

**Regulación y metabolismo evolucionan de forma independiente**

También es un hecho reseñable el que reguladores diferentes, que pertenecen a distintas familias sean capaces de responder a los mismos compuestos y que reguladores similares, de la misma familia o incluso muy parecidos en secuencia no respondan necesariamente a compuestos similares. En definitiva, no parece existir un patrón claro que correlacione homología entre reguladores con similitud entre compuestos. Por otra parte, observamos que operones homólogos, según unos criterios de identidad de secuencia entre genes, no tienen necesariamente reguladores homólogos. Esto parece sugerir que la regulación y el metabolismo evolucionan de forma independiente como ya se sugirió anteriormente llamando a este fenómeno gato negro/gato blanco (Cases y de Lorenzo 2001) en referencia al hecho de que distintos mecanismos regulatorios producen el mismo resultado. La escasez de datos nos impide ser concluyentes en esta afirmación, pero sí encontramos un mayor nivel de conservación entre los genes de los reguladores de operones homólogos cuando el gen del regulador está contiguo al conjunto de genes que regula. En *Escherichia* coli, se ha observado que los reguladores que se encuentran adyacentes a los genes que regulan se encuentran también con más frecuencia adyacentes a sus genes regulados en otros genomas (Korbel *et al.*, 2004). Quizá en biodegradación ocurra lo mismo y una vez que un circuito lógico de regulación se ha ensamblado de forma física su organización genética se conserve porque favorece la adquisición del control de las capacidades catabólicas adquiridas.

**Regulación a nivel de comunidad o las ventajas de compartir**

Si ponemos en común todos los factores que hemos estudiado, nos encontramos ante unos sistemas flexibles pero estables de forma independiente, sin necesidad de un control adicional. La flexibilidad viene dada por el hecho de que se pueden construir a partir de, en principio, cualquier tipo de regulador, tanto represores como activadores. Estos reguladores permiten adquirir respuestas a nuevas señales por a través de la inespecificidad. En los casos en los que los promotores están asociados a sigma 54 se añaden a esta flexibilidad los múltiples mecanismos de este tipo de promotores para integración con la fisiología del huésped (Carmona *et al.*, 1997;Macchi *et al.*, 2003;Van Dien y de Lorenzo, 2003), aunque la mayoría de los promotores están asociados a sigma 70. La estabilidad viene dada gracias al acoplamiento de la regulación y la autorregulación (Wall *et al.*,2005), transferibles por su organización compacta que constituye un módulo funcional que ocupa el mínimo espacio de ADN y se transfiere junto. La evolución independiente de las enzimas y de los reguladores sugiere que los operones catabólicos se han integrado en las bacterias a las que han podido llegar a través de la maquinaria regulatoria disponible en esa bacteria huésped. Existen numerosos sistemas para controlar de integración de la expresión del promotor con la fisiología del huésped. Por ejemplo, en el caso del promotor *Pu* del plásmido TOL existe un modelo dinámico matemático (Van Dien y de Lorenzo, 2003) en el que se muestra un gran número de factores que influyen en la integración con la fisiología de la expresión del promotor: la unión del activador a su sitio de unión al ADN, la unión del factor sigma al núcleo de la ARNpolimerasa, la formación del complejo abierto, la liberación de la maquinaria de transcripción de la región del promotor, el control dependiente de la fase de crecimiento de IHF o la contribución de ppGpp a la selección del factor sigma y a la liberación del promotor. Una combinación de tres de estos factores podría producir el mismo comportamiento: represión durante el crecimiento exponencial y un rápido aumento de la actividad cuando las células entran en la fase estacionaria. Entonces, ¿por qué disponer de mecanismos redundantes? Se ha sugerido que la selección de estos sistemas se podría realizar a nivel de su integración en una comunidad microbiana más que en un individuo (Cases y de Lorenzo, 2005). Estos mecanismos que para la integración en un solo organismo resultan redundantes facilitan la interacción de los promotores se integren en la fisiología

de un mayor número de organismos permitiendo a un mayor número de bacterias realizar una actividad que va a ser beneficiosa para toda la comunidad y por ello puede ser seleccionada a ese nivel: eliminar compuestos tóxicos del medio ambiente. De esta forma, la selección actúa a varios niveles. Los operones catabólicos se pueden adquirir del medio por transformación o conjugación. Una vez adquiridos se puede acoplar su expresión a reguladores disponibles en el organismo huésped que sean poco específicos para que puedan reconocer una señal relacionada con el proceso catabólico que realiza un operón permitiendo que este se exprese cuando sea útil. Si esto ocurre, este sistema regulatorio puede ser seleccionado a favor. Una vez acoplada la respuesta esta se puede refinar e irse seleccionando la señal que coincide con el primer sustrato de la ruta expresada lo que permite una respuesta más rápida(Wall *et al.*, 2004). Si a continuación también integra la represión del gen del regulador el sistema será aún más estable y robusto y de respuesta más rápida (Wall *et al.*, 2004) lo que también puede ser seleccionado a favor. Finalmente, la posición contigua del gen del regulador al operón catabólico regulado y su transcripción divergente también ayuda a obtener una respuesta más rápida y facilitan la regulación coordinada lo que supone una ventaja adaptativa. Por otra parte, para que el operón catabólico adquirido sea seleccionado debe proporcionar alguna ventaja al organismo que o adquiere, ya sea en forma de un complejo enzimático completo o una ruta catabólica de degradación de un compuesto tóxico. Probablemente no sea de mucha utilidad adquirir la capacidad de degradar un compuesto tóxico sólo parcialmente ya que la amenaza de su toxicidad sigue presente en el compuesto intermedio. Quizá por eso se seleccionen con más frecuencia operones largos que codifiquen varios pasos de una ruta catabólica. Es cierto que se puede argumentar que un operón largo es tan transferible como muchos pequeños. Se ha propuesto que la formación de nuevos operones reduce la cantidad de información regulatoria necesaria para especificar patrones óptimos de expresión y que es más probable que se formen nuevos operones que promotores independientes cuando la regulación es compleja. El hecho de que los operones tengan sus secuencias reguladoras más conservadas que los genes transcritos de forma independiente apoya esta hipótesis (Price *et al.*, 2005). Teniendo esto en cuenta, la necesidad de adquirir sitios de unión para los reguladores para un mayor número de genes, así como el derroche de energía que puede suponer tener que expresar un número mayor de reguladores para controlar operones

cortos puede inclinar la balanza hacia la selección de operones largos. Si junto a uno de estos operones largos encontramos un regulador anexo, que como hemos observado es algo frecuente y que puede seleccionarse a favor por otras razones se termina constituyendo un módulo estable que permite la degradación de un compuesto tóxico cuando este aparece en el medio. Esta organización compacta del módulo también permite su adquisición con la regulación incluida y su establecimiento en un nuevo huésped que puede integrar la expresión a su fisiología por medio de uno de los varios mecanismos disponibles. De este modo, se van modelando unos circuitos de regulación no sólo por las ventajas que puedan ofrecer a un organismo concreto si no por ayudar a la supervivencia de toda una comunidad que así puede resistir ante una amenaza tóxica por la acción coordinada de los operones catabólicos en distintos organismo que forman conjuntamente 'estimulones', conjuntos de genes que se expresan en respuesta a un estímulo externo, que se expresan de forma conjunta.

Resumiendo, los mecanismos de regulación de los operones catabólicos implicados en biodegradación pueden tener diferentes orígenes evolutivos pero presentan unas características comunes que los hacen especialmente flexibles en su evolución y estables una vez formados comparados con los de un organismo modelo como *Escherichia coli*. La organización genética de los componentes de los circuitos de regulación es compacta y favorece la co-regulación y su adquisición a través de elementos móviles, como plásmidos o transposones, o en procesos de transformación. La poca especificidad de los reguladores puede ser una forma de integrar nuevas señales. Los operones catabólicos parecen evolucionar de forma independiente de los mecanismos de regulación lo que sugiere la idea de que la flexibilidad en los mecanismos de regulación permite la integración de los operones que se puedan haber adquirido utilizando los mecanismos de regulación disponibles en el microorganismo huésped. Esta posibilidad de integrar nuevas capacidades puede permitir a los diferentes microorganismos aumentar su rango de degradación de compuestos tóxicos. Sugerimos la existencia de un ciclo que influye en los sistemas de regulación a, al menos, 3 niveles: selección de los mecanismos de regulación a nivel de circuito regulatorio, modelado de la organización genética por la transferencia entre bacterias e integración de nuevas capacidades biodegradativas por la flexibilidad de la especificidad de los reguladores y promotores.

# CONCLUSIONES

1. La proporción de activadores en los sistemas de biodegradación es mucho mayor que la de cualquier otro tipo de regulador y no está causada por un efecto fundador, y, en consonancia con la mayor proporción de activadores, hay una mayor proporción de promotores catabólicos inducibles controlados por activadores.
2. La diferente acción, activación o represión, ejercida por los reguladores, en función de la distancia a la que se unen con respecto al inicio de transcripción, permite inducir los promotores catabólicos y reprimir el gen del regulador, lo que hace que los circuitos sean más estables, robustos y tengan una mayor capacidad de respuesta.
3. Los reguladores específicos en biodegradación controlan más genes por medio de menos promotores que *E. coli* debido a que los operones son más largos y la regulación de genes y promotores es más sencilla.
4. En biodegradación existe una mayor proporción de promotores sigma 54 que en *E. coli* que no está causada por un efecto fundador, lo que podría estar relacionado con la capacidad de sigma 54 de integrar los promotores con la fisiología del organismo
5. Los complejos enzimáticos casi siempre están agrupados en operones, o en grupos de genes contiguos en su defecto, y ordenados de manera consecutiva cuando hacen reacciones consecutivas.
6.Los genes catabólicos y el gen del regulador que controla su expresión se encuentran frecuentemente adyacentes lo que facilita su co-regulación, al solapar las zonas de regulación de sus promotores, y su movilidad por transferencia horizontal de genes como una unidad funcional completa.
7.La regulación se ha acoplado en numerosas ocasiones a los operones catabólicos pero tiene una tendencia a estar más conservada cuando el gen del regulador está adyacente a sus genes regulados, lo que podría estar relacionado con la capacidad de adquirir un circuito funcional completo en un evento de transferencia horizontal de genes.
8. El orden de los genes en los operones catabólicos favorece la transferencia de complejos enzimáticos y rutas catabólicas.
9. Los reguladores son más promiscuos en los sistemas de biodegradación y, tanto su amplio rango de reconocimiento de señales como la gran cantidad de inducción gratuita de sus operones regulados, apoyan la teoría del 'ruido regulatorio', que postula que para

la adquisición de nuevas señales los sistemas de regulación parten de una situación de poca especificidad.

# BIBLIOGRAFÍA


Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990): Basic local alignment search tool. *J Mol Biol* **215**, 403-10.

Arias-Barrau, E., Olivera, E. R., Luengo, J. M., Fernandez, C., Galan, B., Garcia, J. L., Diaz, E., and Minambres, B. (2004): The homogentisate pathway: a central catabolic pathway involved in the degradation of L-phenylalanine, L-tyrosine, and 3-hydroxyphenylacetate in Pseudomonas putida. *J Bacteriol* **186**, 5062-77.

Baldi, P., Benz, R. W., Hirschberg, D. S., and Swamidass, S. J. (2007): Lossless compression of chemical fingerprints using integer entropy codes improves storage and retrieval. *J Chem Inf Model* **47**, 2098-109.

Baldi, P., and Benz, R. W. (2008): BLASTing small molecules--statistics and extreme statistics of chemical similarity scores. *Bioinformatics* **24**, i357-65.

Barabasi, A. L., and Albert, R. (1999): Emergence of scaling in random networks. *Science* **286**, 509-12.

Barabasi, A. L., and Bonabeau, E. (2003): Scale-free networks. *Sci Am* **288**, 60-9.

Barabasi, A. L., and Oltvai, Z. N. (2004): Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101-13.

Barragan, M. J., Blazquez, B., Zamarro, M. T., Mancheno, J. M., Garcia, J. L., Diaz, E., and Carmona, M. (2005): BzdR, a repressor that controls the anaerobic catabolism of benzoate in Azoarcus sp. CIB, is the first member of a new subfamily of transcriptional regulators. *J Biol Chem* **280**, 10683-94.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004): The Pfam protein families database. *Nucleic Acids Res* **32**, D138-41.

Becskei, A., and Serrano, L. (2000): Engineering stability in gene networks by autoregulation. *Nature* **405**, 590-3.

Bertoni, G., Fujita, N., Ishihama, A., and de Lorenzo, V. (1998): Active recruitment of sigma54-RNA polymerase to the Pu promoter of Pseudomonas putida: role of IHF and alphaCTD. *Embo J* **17**, 5120-8.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003): The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365-70.

Brinkrolf, K., Brune, I., and Tauch, A. (2006): Transcriptional regulation of catabolic pathways for aromatic compounds in Corynebacterium glutamicum. *Genet Mol Res* **5**, 773-89.

Brzostowicz, P. C., Reams, A. B., Clark, T. J., and Neidle, E. L. (2003): Transcriptional cross-regulation of the catechol and protocatechuate branches of the beta-ketoadipate pathway contributes to carbon source-dependent expression of the Acinetobacter sp. strain ADP1 pobA gene. *Appl Environ Microbiol* **69**, 1598-606.

Carbajosa, G., Trigo, A., Valencia, A., and Cases, I. (2009): Bionemo: molecular information on biodegradation metabolism. *Nucleic Acids Res* **37**, D598-602.

Carmona, M., Claverie-Martin, F., and Magasanik, B. (1997): DNA bending and the initiation of transcription at sigma54-dependent bacterial promoters. *Proc Natl Acad Sci U S A* **94**, 9568-72.

Carmona, M., Rodriguez, M. J., Martinez-Costa, O., and De Lorenzo, V. (2000): In vivo and in vitro effects of (p)ppGpp on the sigma(54) promoter Pu of the TOL plasmid of Pseudomonas putida. *J Bacteriol* **182**, 4711-8.

Cases, I., and de Lorenzo, V. (1998): Expression systems and physiological control of promoter activity in bacteria. *Curr Opin Microbiol* **1**, 303-10.

Cases, I., and de Lorenzo, V. (2001): The black cat/white cat principle of signal integration in bacterial promoters. *Embo J* **20**, 1-11.

Cases, I., and de Lorenzo, V. (2005a): Genetically modified organisms for the environment: stories of success and failure and what we have learned from them. *Int Microbiol* **8**, 213-22.

Cases, I., and de Lorenzo, V. (2005b): Promoters in the environment: transcriptional regulation in its natural context. *Nat Rev Microbiol* **3**, 105-18.

Clark, T. J., Phillips, R. S., Bundy, B. M., Momany, C., and Neidle, E. L. (2004): Benzoate decreases the binding of cis,cis-muconate to the BenM regulator despite the synergistic effect of both compounds on transcriptional activation. *J Bacteriol* **186**, 1200-4.

Collado-Vides, J., Magasanik, B., and Gralla, J. D. (1991): Control site location and transcriptional regulation in Escherichia coli. *Microbiol Rev* **55**, 371-94.

Coschigano, P. W., Wehrman, T. S., and Young, L. Y. (1998): Identification and analysis of genes involved in anaerobic toluene metabolism by strain T1: putative role of a glycine free radical. *Appl Environ Microbiol* **64**, 1650-6.

Chugani, S. A., Parsek, M. R., and Chakrabarty, A. M. (1998): Transcriptional repression mediated by LysR-type regulator CatR bound at multiple binding sites. *J Bacteriol* **180**, 2367-72.

de Daruvar, A., Collado-Vides, J., and Valencia, A. (2002): Analysis of the cellular functions of Escherichia coli operons and their conservation in Bacillus subtilis. *J Mol Evol* **55**, 211-21.

de Lorenzo, V., Herrero, M., Metzke, M., and Timmis, K. N. (1991): An upstream XylR- and IHF-induced nucleoprotein complex regulates the sigma 54-dependent Pu promoter of TOL plasmid. *Embo J* **10**, 1159-67.

de Lorenzo, V., and Perez-Martin, J. (1996): Regulatory noise in prokaryotic promoters: how bacteria learn to respond to novel environmental signals. *Mol Microbiol* **19**, 1177-84.

Devos, D., and Valencia, A. (2000): Practical limits of function prediction. *Proteins* **41**, 98-107.

Diaz, E., and Prieto, M. A. (2000): Bacterial promoters triggering biodegradation of aromatic pollutants. *Curr Opin Biotechnol* **11**, 467-75.

Díaz, E. (2004): Bacterial degradation of aromatic pollutants: a paradigm of metabolic versatility. *Int Microbiol* **3**, 173-180.

Durante-Rodriguez, G., Zamarro, M. T., Garcia, J. L., Diaz, E., and Carmona, M. (2008): New insights into the BzdR-mediated transcriptional regulation of the anaerobic catabolism of benzoate in Azoarcus sp. CIB. *Microbiology* **154**, 306-16.

Egland, P. G., and Harwood, C. S. (1999): BadR, a new MarR family member, regulates anaerobic benzoate degradation by Rhodopseudomonas palustris in concert with AadR, an Fnr family member. *J Bacteriol* **181**, 2102-9.

Ellis, L. B., Roe, D., and Wackett, L. P. (2006): The University of Minnesota Biocatalysis/ Biodegradation Database: the first decade. *Nucleic Acids Res* **34**, D517-21.

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002): An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-84.

Ermolaeva, M. D., White, O., and Salzberg, S. L. (2001): Prediction of operons in microbial genomes. *Nucleic Acids Res* **29**, 1216-21.

Ezezika, O. C., Haddad, S., Clark, T. J., Neidle, E. L., and Momany, C. (2007): Distinct effector-binding sites enable synergistic transcriptional activation by BenM, a LysR-type regulator. *J Mol Biol* **367**, 616-29.

Fujihara, H., Yoshida, H., Matsunaga, T., Goto, M., and Furukawa, K. (2006): Cross-regulation of biphenyl- and salicylate-catabolic genes by two regulatory systems in Pseudomonas pseudoalcaligenes KF707. *J Bacteriol* **188**, 4690-7.

Furukawa, K., Suenaga, H., and Goto, M. (2004): Biphenyl dioxygenases: functional versatilities and directed evolution. *J Bacteriol* **186**, 5189-96.

Furukawa, K., and Fujihara, H. (2008): Microbial degradation of polychlorinated biphenyls: biochemical and molecular features. *J Biosci Bioeng* **105**, 433-49.

Galan, B., Kolb, A., Sanz, J. M., Garcia, J. L., and Prieto, M. A. (2003): Molecular determinants of the hpa regulatory system of Escherichia coli: the HpaR repressor. *Nucleic Acids Res* **31**, 6598-609.

Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. (2002): Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* **19**, 2226-38.

Goosen, N., and van de Putte, P. (1995): The regulation of transcription initiation by integration host factor. *Mol Microbiol* **16**, 1-7.

Gury, J., Barthelmebs, L., Tran, N. P., Divies, C., and Cavin, J. F. (2004): Cloning, deletion, and characterization of PadR, the transcriptional repressor of the phenolic acid decarboxylase-encoding padA gene of Lactobacillus plantarum. *Appl Environ Microbiol* **70**, 2146-53.

Hacker, J., Blum-Oehler, G., Muhldorfer, I., and Tschape, H. (1997): Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol Microbiol* **23**, 1089-97.

Haranczyk, M., and Holliday, J. (2008): Comparison of similarity coefficients for clustering and compound selection. *J Chem Inf Model* **48**, 498-508.

Henkin, T. M., and Yanofsky, C. (2002): Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions. *Bioessays* **24**, 700-7.

Hershberg, R., Yeger-Lotem, E., and Margalit, H. (2005): Chromosomal organization is shaped by the transcription regulatory network. *Trends Genet* **21**, 138-42.

Hlavacek, W. S., and Savageau, M. A. (1995): Subunit structure of regulator proteins influences the design of gene circuitry: analysis of perfectly coupled and completely uncoupled circuits. *J Mol Biol* **248**, 739-55.

Hlavacek, W. S., and Savageau, M. A. (1996): Rules for coupled expression of regulator and effector genes in inducible circuits. *J Mol Biol* **255**, 121-39.

Hlavacek, W. S., and Savageau, M. A. (1997): Completely uncoupled and perfectly coupled gene expression in repressible systems. *J Mol Biol* **266**, 538-58.

Houghton, J. E., Brown, T. M., Appel, A. J., Hughes, E. J., and Ornston, L. N. (1995): Discontinuities in the evolution of Pseudomonas putida cat genes. *J Bacteriol* **177**, 401-12.

Holliday, J. D., Hu, C. Y., and Willett, P. (2002): Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Comb Chem High Throughput Screen* **5**, 155-66.

Itoh, T., Takemoto, K., Mori, H., and Gojobori, T. (1999): Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol Biol Evol* **16**, 332-46.

Jacob, F., Perrin, D., Sanchez, C., and Monod, J. (1960): [Operon: a group of genes with the expression coordinated by an operator.]. *C R Hebd Seances Acad Sci* **250**, 1727-9.

Jacob, F., and Monod, J. (1961a): Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**, 318-56.

Janga, S. C., and Collado-Vides, J. (2007a): Structure and evolution of gene regulatory networks in microbial genomes. *Res Microbiol* **158**, 787-94.

Janga, S. C., Salgado, H., Collado-Vides, J., and Martinez-Antonio, A. (2007b): Internal versus external effector and transcription factor gene pairs differ in their relative chromosomal position in Escherichia coli. *J Mol Biol* **368**, 263-72.

Janssen, D. B., Dinkla, I. J., Poelarends, G. J., and Terpstra, P. (2005): Bacterial degradation of xenobiotic compounds: evolution and distribution of novel enzyme activities. *Environ Microbiol* **7**, 1868-82.

Johnson, G. R., Jain, R. K., and Spain, J. C. (2002): Origins of the 2,4-dinitrotoluene pathway. *J Bacteriol* **184**, 4219-32.

Keseler, I. M., Bonavides-Martinez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R. P., Johnson, D. A., Krummenacker, M., Nolan, L. M., Paley, S., Paulsen, I. T., Peralta-Gil, M., Santos-Zavaleta, A., Shearer, A. G., and Karp, P. D. (2009): EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Res* **37**, D464-70.

Kitagawa, W., Takami, S., Miyauchi, K., Masai, E., Kamagata, Y., Tiedje, J. M., and Fukuda, M. (2002): Novel 2,4-dichlorophenoxyacetic acid degradation genes from oligotrophic Bradyrhizobium sp. strain HW13 isolated from a pristine environment. *J Bacteriol* **184**, 509-18.

Korbel, J. O., Jensen, L. J., von Mering, C., and Bork, P. (2004): Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat Biotechnol* **22**, 911-7.

Kulakov, L. A., Chen, S., Allen, C. C., and Larkin, M. J. (2005): Web-type evolution of rhodococcus gene clusters associated with utilization of naphthalene. *Appl Environ Microbiol* **71**, 1754-64.

Lathe, W. C., 3rd, Snel, B., and Bork, P. (2000): Gene context conservation of a higher order than operons. *Trends Biochem Sci* **25**, 474-9.

Lawrence, J. G., and Roth, J. R. (1996): Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**, 1843-60.

Leelakriangsak, M., Huyen, N. T., Towe, S., van Duy, N., Becher, D., Hecker, M., Antelmann, H., and Zuber, P. (2008): Regulation of quinone detoxification by the thiol stress sensing DUF24/MarR-like repressor, YodB in Bacillus subtilis. *Mol Microbiol* **67**, 1108-24.

Leoni, L., Rampioni, G , Zennaro, E (2007): Styrene, an Unpalatable Substrate with Complex Regulatory Networks, pp. 59-87. In J.-L. R. a. A. Filloux (Ed.): *Pseudomonas*, Springer Netherlands.

Macchi, R., Montesissa, L., Murakami, K., Ishihama, A., De Lorenzo, V., and Bertoni, G. (2003): Recruitment of sigma54-RNA polymerase to the Pu promoter of Pseudomonas putida through integration host factor-mediated positioning switch of alpha subunit carboxyl-terminal domain on an UP-like element. *J Biol Chem* **278**, 27695-702.

Madan Babu, M., and Teichmann, S. A. (2003): Functional determinants of transcription factors in Escherichia coli: protein families and binding sites. *Trends Genet* **19**, 75-9.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999): Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751-3.

Martinez-Antonio, A., and Collado-Vides, J. (2003): Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol* **6**, 482-9.

McAdams, H. H., Srinivasan, B., and Arkin, A. P. (2004): The evolution of genetic regulatory systems in bacteria. *Nat Rev Genet* **5**, 169-78.

Morales, G., Linares, J. F., Beloso, A., Albar, J. P., Martinez, J. L., and Rojo, F. (2004): The Pseudomonas putida Crc global regulator controls the expression of genes from several chromosomal catabolic pathways for aromatic compounds. *J Bacteriol* **186**, 1337-44.

Moreno, R., Ruiz-Manzano, A., Yuste, L., and Rojo, F. (2007): The Pseudomonas putida Crc global regulator is an RNA binding protein that inhibits translation of the AlkS transcriptional regulator. *Mol Microbiol* **64**, 665-75.

Muller, T. A., Werlen, C., Spain, J., and Van Der Meer, J. R. (2003): Evolution of a chlorobenzene degradative pathway among bacteria in a contaminated groundwater mediated by a genomic island in Ralstonia. *Environ Microbiol* **5**, 163-73.

Mushegian, A. R., and Koonin, E. V. (1996): Gene order is not conserved in bacterial evolution. *Trends Genet* **12**, 289-90.

Olivera, E. R., Minambres, B., Garcia, B., Muniz, C., Moreno, M. A., Ferrandez, A., Diaz, E., Garcia, J. L., and Luengo, J. M. (1998): Molecular characterization of the phenylacetic acid catabolic pathway in Pseudomonas putida U: the phenylacetyl-CoA catabolon. *Proc Natl Acad Sci U S A* **95**, 6419-24.

Paget, M. S., and Helmann, J. D. (2003): The sigma70 family of sigma factors. *Genome Biol* **4**, 203.

Parales, R. E., Bruce, N. C., Schmid, A., and Wackett, L. P. (2002): Biodegradation, biotransformation, and biocatalysis (b3). *Appl Environ Microbiol* **68**, 4699-709.

Park, H. S., and Kim, H. S. (2001): Genetic and structural organization of the aminophenol catabolic operon and its implication for evolutionary process. *J Bacteriol* **183**, 5074-81.

Park, H. H., Lee, H. Y., Lim, W. K., and Shin, H. J. (2005): NahR: effects of replacements at Asn 169 and Arg 248 on promoter binding and inducer recognition. *Arch Biochem Biophys* **434**, 67-74.

Parsek, M. R., Ye, R. W., Pun, P., and Chakrabarty, A. M. (1994): Critical nucleotides in the interaction of a LysR-type regulator with its target promoter region. catBC promoter activation by CatR. *J Biol Chem* **269**, 11279-84.

Price, M. N., Huang, K. H., Arkin, A. P., and Alm, E. J. (2005): Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res* **15**, 809-19.

Pazos, F., Valencia, A., and De Lorenzo, V. (2003): The organization of the microbial biodegradation network from a systems-biology perspective. *EMBO Rep* **4**, 994-9.

Pazos, F., Guijas, D., Valencia, A., and De Lorenzo, V. (2005): MetaRouter: bioinformatics for bioremediation. *Nucleic Acids Res* **33**, D588-92.

Perez-Martin, J., and De Lorenzo, V. (1995): Integration host factor suppresses promiscuous activation of the sigma 54-dependent promoter Pu of Pseudomonas putida. *Proc Natl Acad Sci U S A* **92**, 7277-81.

PostgreSQL www.postgresql.org. (Accesed May 21, 2007).

Providenti, M. A., and Wyndham, R. C. (2001): Identification and functional characterization of CbaR, a MarR-like modulator of the cbaABC-encoded chlorobenzoate catabolism pathway. *Appl Environ Microbiol* **67**, 3530-41.

Ramos, J. L., Stolz, A., Reineke, W., and Timmis, K. N. (1986): Altered effector specificities in regulators of gene expression: TOL plasmid xylS mutants and their use to engineer expansion of the range of aromatics degraded by bacteria. *Proc Natl Acad Sci U S A* **83**, 8467-71.

Ramos, J. L., Marques, S., and Timmis, K. N. (1997): Transcriptional control of the Pseudomonas TOL plasmid catabolic operons is achieved through an interplay of host factors and plasmid-encoded regulators. *Annu Rev Microbiol* **51**, 341-73.

Reitzer, L., and Schneider, B. L. (2001): Metabolic context and possible physiological themes of sigma(54)-dependent genes in Escherichia coli. Microbiol Mol Biol Rev 65, 422-44, table of contents.

Rhee, K. Y., Opel, M., Ito, E., Hung, S., Arfin, S. M., and Hatfield, G. W. (1999): Transcriptional coupling between the divergent promoters of a prototypic LysR-type regulatory system, the ilvYC operon of Escherichia coli. *Proc Natl Acad Sci U S A* **96**, 14294-9.

Rogozin, I. B., Makarova, K. S., Murvai, J., Czabarka, E., Wolf, Y. I., Tatusov, R. L., Szekely, L. A., and Koonin, E. V. (2002): Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res* **30**, 2212-23.

Rocha, E. P. (2006): Inference and analysis of the relative stability of bacterial chromosomes. Mol Biol Evol 23, 513-22.

Rojo, F., Pieper, D. H., Engesser, K. H., Knackmuss, H. J., and Timmis, K. N. (1987): Assemblage of ortho cleavage route for simultaneous degradation of chloro- and methylaromatics. *Science* **238**, 1395-8.

Rogozin, I. B., Makarova, K. S., Murvai, J., Czabarka, E., Wolf, Y. I., Tatusov, R. L., Szekely, L. A., and Koonin, E. V. (2002): Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res* **30**, 2212-23.

Romero-Steiner, S., Parales, R. E., Harwood, C. S., and Houghton, J. E. (1994): Characterization of the pcaR regulatory gene from Pseudomonas putida, which is required for the complete degradation of p-hydroxybenzoate. *J Bacteriol* **176**, 5771-9.

Rosenfeld, N., Elowitz, M. B., and Alon, U. (2002): Negative autoregulation speeds the response times of transcription networks. *J Mol Biol* **323**, 785-93.

Ross Ihaka and Robert Gentleman. R (1996): A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**(3),299-314

Ruiz-Manzano, A., Yuste, L., and Rojo, F. (2005): Levels and activity of the Pseudomonas putida global regulatory protein Crc vary according to growth conditions. *J Bacteriol* **187**, 3678-86.

Salgado, H., Moreno-Hagelsieb, G., Smith, T. F., and Collado-Vides, J. (2000): Operons in Escherichia coli: genomic analyses and predictions. *Proc Natl Acad Sci U S A* **97**, 6652-7.

Savageau, M. A. (1974): Comparison of classical and autogenous systems of regulation in inducible operons. *Nature* **252**, 546-9.

Savageau, M. A. (1975): Significance of autogenously regulated and constitutive synthesis of regulatory proteins in repressible biosynthetic systems. *Nature* **258**, 208-14.

Savageau, M. A. (1977): Design of molecular control mechanisms and the demand for gene expression. *Proc Natl Acad Sci U S A* **74**, 5647-51.

Selifonova, O. V., and Eaton, R. W. (1996): Use of an ipb-lux Fusion To Study Regulation of the Isopropylbenzene Catabolism Operon of Pseudomonas putida RE204 and To Detect Hydrophobic Pollutants in the Environment. *Appl Environ Microbiol* **62**, 778-783.

Shinar, G., Dekel, E., Tlusty, T., and Alon, U. (2006): Rules for biological regulation based on error minimization. *Proc Natl Acad Sci U S A* **103**, 3999-4004.

Shingler, V. (2003): Integrated regulation in response to aromatic compounds: from signal sensing to attractive behaviour. *Environ Microbiol* **5**, 1226-41.

Smith, M., Kunin, V., Goldovsky, L., Enright, A. J., and Ouzounis, C. A. (2005): MagicMatch--cross-referencing sequence identifiers across databases. *Bioinformatics* **21**, 3429-30.

Snel, B., Lehmann, G., Bork, P., and Huynen, M. A. (2000): STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* **28**, 3442-4.

Solera, D., Arenghi, F. L., Woelk, T., Galli, E., and Barbieri, P. (2004): TouR-mediated effector-independent growth phase-dependent activation of the sigma54 Ptou promoter of Pseudomonas stutzeri OX1. *J Bacteriol* **186**, 7353-63.

Sorensen, S. J., Bailey, M., Hansen, L. H., Kroer, N., and Wuertz, S. (2005): Studying plasmid horizontal transfer in situ: a critical review. *Nat Rev Microbiol* **3**, 700-10.

Springael, D., and Top, E. M. (2004): Horizontal gene transfer and microbial adaptation to xenobiotics: new types of mobile genetic elements and lessons from ecological studies. *Trends Microbiol* **12**, 53-8.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. (2002): The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12**, 1611-8.

Szalewska-Palasz, A., Johansson, L. U., Bernardo, L. M., Skarfstad, E., Stec, E., Brannstrom, K., and Shingler, V. (2007): Properties of RNA polymerase bypass mutants: implications for the role of ppGpp and its co-factor DksA in controlling transcription dependent on sigma54. *J Biol Chem* **282**, 18046-56.

Sze, C. C., Laurie, A. D., and Shingler, V. (2001): In vivo and in vitro effects of integration host factor at the DmpR-regulated sigma(54)-dependent Po promoter. *J Bacteriol* **183**, 2842-51.

Tamames, J. (2001): Evolution of gene order conservation in prokaryotes. *Genome Biol* **2**, RESEARCH0020.

Tamames, J., Gonzalez-Moreno, M., Mingorance, J., Valencia, A., and Vicente, M. (2001): Bringing gene order into bacterial shape. *Trends Genet* **17**, 124-6.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. (2001): The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29**, 22-8.

Thanaraj, T. A., and Argos, P. (1996): Ribosome-mediated translational pause and protein domain organization. *Protein Sci* **5**, 1594-612.

Thanbichler, M., Iniesta, A. A., and Shapiro, L. (2007): A comprehensive set of plasmids for vanillate- and xylose-inducible gene expression in Caulobacter crescentus. *Nucleic Acids Res* **35**, e137.

The Source for Perl www.perl.com. (Accesed May 17, 2007.).

Thieffry, D., Huerta, A. M., Perez-Rueda, E., and Collado-Vides, J. (1998): From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. *Bioessays* **20**, 433-40.

Thomas, C. M., and Nielsen, K. M. (2005): Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* **3**, 711-21.

Top, E. M., and Springael, D. (2003): The role of mobile genetic elements in bacterial adaptation to xenobiotic organic compounds. *Curr Opin Biotechnol* **14**, 262-9.

Tropel, D., and van der Meer, J. R. (2004): Bacterial transcriptional regulators for degradation pathways of aromatic compounds. *Microbiol Mol Biol Rev* **68**, 474-500.

Townsend, J. P., Nielsen, K. M., Fisher, D. S., and Hartl, D. L. (2003): Horizontal acquisition of divergent chromosomal DNA in bacteria: effects of mutator phenotypes. *Genetics* **164**, 13-21. Tropel, D., and van der Meer, J. R. (2004): Bacterial transcriptional regulators for degradation pathways of aromatic compounds. *Microbiol Mol Biol Rev* **68**, 474-500.

van Beilen, J. B., Marin, M. M., Smits, T. H., Rothlisberger, M., Franchini, A. G., Witholt, B., and Rojo, F. (2004): Characterization of two alkane hydroxylase genes from the marine hydrocarbonoclastic bacterium Alcanivorax borkumensis. *Environ Microbiol* **6**, 264-73.

van der Meer, J. R., de Vos, W. M., Harayama, S., and Zehnder, A. J. (1992): Molecular mechanisms of genetic adaptation to xenobiotic compounds. *Microbiol Rev* **56**, 677-94.

van der Meer, J. R., and Sentchilo, V. (2003): Genomic islands and the evolution of catabolic pathways in bacteria. *Curr Opin Biotechnol* **14**, 248-54.

Van Dien, S. J., and de Lorenzo, V. (2003): Deciphering environmental signal integration in sigma54-dependent promoters with a simple mathematical model. *J Theor Biol* **224**, 437-49.

Velazquez, F., Parro, V., and de Lorenzo, V. (2005): Inferring the genetic network of m-xylene metabolism through expression profiling of the xyl genes of Pseudomonas putida mt-2. *Mol Microbiol* **57**, 1557-69.

Velazquez, F., de Lorenzo, V., and Valls, M. (2006): The m-xylene biodegradation capacity of Pseudomonas putida mt-2 is submitted to adaptation to abiotic stresses: evidence from expression profiling of xyl genes. *Environ Microbiol* **8**, 591-602.

Wall, M. E., Hlavacek, W. S., and Savageau, M. A. (2003): Design principles for regulator gene expression in a repressible gene circuit. *J Mol Biol* **332**, 861-76.

Wall, M. E., Hlavacek, W. S., and Savageau, M. A. (2004): Design of gene circuits: lessons from bacteria. *Nat Rev Genet* **5**, 34-42.

Warren, P. B., and ten Wolde, P. R. (2004): Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of Escherichia coli. *J Mol Biol* **342**, 1379-90.

Watanabe, H., Mori, H., Itoh, T., and Gojobori, T. (1997): Genome plasticity as a paradigm of eubacteria evolution. *J Mol Evol* **44 Suppl 1**, S57-64

Watanabe, T., Fujihara, H., and Furukawa, K. (2003): Characterization of the second LysR-type regulator in the biphenyl-catabolic gene cluster of Pseudomonas pseudoalcaligenes KF707. *J Bacteriol* **185**, 3575-82.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Ostell, J., Pruitt, K. D., Schuler, G. D., Shumway, M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2008): Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **36**, D13-21.

Woese, C. R. (2004): A new biology for a new century. *Microbiol Mol Biol Rev* **68**, 173-86.

Xu, H., and Hoover, T. R. (2001): Transcriptional regulation at a distance in bacteria. *Curr Opin Microbiol* **4**, 138-44.

# APÉNDICE

Bionemo: molecular information on biodegradation metabolism

Reunido el tribunal que suscribe en el día de la fecha, acordó calificar la presente Tesis doctoral con SOBRESALIENTE CUM LAUDE

MADRID 16 FEB 2009

VICENTE DE LUCENZO

José Bergua

FLORENCIO PAÑOS

Fdo. Eduardo Santero

JAVIER TAMAMES